DOCUMENT RESUME

ED 363 096 FL 021 487

AUTHOR	Huhta, Ari, Ed.; And Others
TITLE	Language Testing: New Openings.
INSTITUTION	Jyvaskyla Univ. (Finland). Inst. for Educational Research.
REPORT NO	ISBN-951-34-0059-X
PUB DATE	93
NOTE	252p.
PUB TYPE	Collected Works - General (020) -- Books (010)
EDRS PRICE	MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS	Cloze Procedure; Criterion Referenced Tests; Difficulty Level; Educational Trends; English (Second Language); Foreign Countries; Interference (Language); *Language Planning; Language Proficiency; *Language Tests; Oral Language; Personality Traits; Politics of Education; *Public Policy; Reading Comprehension; *Second Languages; Set Theory; Standardized Tests; Student Characteristics; Test Bias; *Test Construction; *Testing; Transfer of Training; Trend Analysis; Writing Evaluation
IDENTIFIERS	Finland; Iconicity; Test of English as a Foreign Language

ABSTRACT

A collection of articles on second-language testing includes: "The State of Language Testing in the 1990s" (J. Charles Alderson); "The Exercise of Power and Control in the Rhetorics of Testing" (Elana Shohamy); "Institutional and Political Aspects of Language Testing" (Bernard Spolsky); "Language Planning Policy and Development of FL Proficiency in Finland" (Sauli Takala); "Language Policy and Language Testing" (Ingrid Wijgh); "Criterion-Referenced Language Test Development: A Prolegomenon" (Fred Davidson, Brian Lynch); "A Method for Improving Tasks on Performance-Based Assessments through Field Testing" (Dorry Mann Kenyon, Charles W. Stansfield); "Fuzzy Set Theory Application in Criterion-Referenced Testing" (Felly Stoyanova); "Personality Characteristics as a Potential Source of Language Test Bias" (Vivien Berry); "Relationships between TOEFL Subsections, Speaking and Language Learning Background" (Harry L. Gradman, Edith Hanania); "Recent Developments in National Examinations in Finland" (Ari Huhta, Kari Sajavaara, Sauli Takala); "Tests for Measuring the Understanding of Iconical Tropes" (Christo Kaftandiiev, Felly Stoyanova); "Rating the Quality of Student Writing--Findings from the IEA Study of Achievement in Written Composition" (Rainer H. Lehmann); "A Model of ESL Reading Comprehension Difficulty" (Kyle Perkins, Sheila L. Brutten); "Question Level Bias in Cloze Questions--an L1 Transfer Effect" (Alastair Pollitt, Lynda Taylor); and "Proficiency Tests in Contemporary Foreign Language Syllabus" (Diana Rumpite, A. Rotanova, L. Ilyanskaya). (MSE)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

LANGUAGE TESTING: NEW OPENINGS

Ari Huhta – Kari Sajavaara – Sauli Takala (Eds.)

ED 363 096

"PERMISSION TO REPRODUCE THIS MATERIAL IN OTHER THAN PAPER COPY HAS BEEN GRANTED BY

Sauli Takala

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

FL021487

BEST COPY AVAILABLE

Language Testing: New Openings

Ari Huhta, Kari Sajavaara and Sauli Takala
Editors

Institute for Educational Research
University of Jyväskylä 1993

3

© Authors and Institute for Educational Research

ISBN 951-34-0059-X

Printed by Oy Sisäsuomi, Jyväskylä, Finland

1993

TABLE OF CONTENTS

INTRODUCTION		v
Alderson, J.Charles	The State of Language Testing in the 1990s.	1

Section I
Language Testing Policy

Shohamy, Elana	The Exercise of Power and Control in the Rhetorics of Testing.	23
Spolsky, Bernard	Institutional and Political Aspects of Language Testing.	39
Takala, Sauli	Language Planning Policy and Development of FL Proficiency in Finland.	46
Wijgh, Ingrid	Language Policy and Language Testing.	55

Section II
Methodological Aspects of Language Testing

Davidson, Fred and Brian Lynch	Criterion-Referenced Language Test Development: A Prolegomenon.	73
Kenyon, Dorry Mann and Charles W. Stansfield	A Method for Improving Tasks on Performance-Based Assessments through Field Testing.	90
Stoyanova, Felly	Fuzzy Set Theory Application in Criterion-Referenced Testing.	103

Section III
Empirical Investigations and Test Development

Berry, Vivien	Personality Characteristics as a Potential Source of Language Test Bias.	115
Gradman, Harry L. and Edith Hanania	Relationships between TOEFL Subsections, Speaking and Language Learning Background.	125
Huhta, Ari, Kari Sajavaara and Sauli Takala	Recent Developments in National Examinations in Finland.	136
Kaftandiiev, Christo and Felly Stoyanova	Tests for Measuring the Understanding of Iconical Tropes.	160
Lehmann, Rainer H.	Rating the Quality of Student Writing — Findings from the IEA Study of Achievement in Written Composition.	186
Perkins, Kyle and Sheila L.Brutten	A Model of ESL Reading Comprehension Difficulty.	205
Pollitt, Alastair and Lynda Taylor	Question Level Bias in Cloze Questions — an L1 Transfer Effect.	219
Rumpite, Diana, A. Rotanova and L. Ilyanskaya	Proficiency Tests in Contemporary Foreign Language Syllabus.	237
APPENDIX	**Authors' addresses**	249

INTRODUCTION

This book is a collection of papers presented at a conference organized by the University of Jyväskylä language testing team in August 1991.

The conference was planned as a broadly based forum for reviewing the state of language testing in Europe and for charting current practices in language testing in both foreign/second languages and the mother tongue. The organizers expressed a wish to promote integration between L1 and L2 researchers. An attempt was also made to acknowledge the linguistic diversity of Europe by inviting presentations in major languages other than English. Presentations were specifically called for in the testing of oral proficiency, reading and writing research and language testing, professionally/vocationally oriented language tests, new technologies in testing, and national or equivalent language examinations.

The organizers of the conference wanted to bring together language testing experts from the East and the West in a situation where the circumstances had become favorable for such a get-together. We felt that, for geographical and other reasons, Finland was a suitable venue for eastern and western practitioners to meet and establish new contacts.

In practice, participation from the East did not materialize to the extent planned. Yet the conference proved to be an opening of a new dialogue. Contributions by the eastern participants show that interesting work on language testing has been carried out in the eastern European countries and that it deserves to be better known by the testing community.

Another opening was thematic, not geographical. The conference was also planned to take a look at the future as well. Charles Alderson was asked to give a keynote speech, present a personal assessment of the state of testing and outline some of the directions worthy of the attention of the testing community. In his paper he identifies several areas in which research would be useful.

The third opening was towards the practitioners: after the conference, several of the international participants participated in a seminar organized in Jyväskylä for Finnish language teachers. They contributed to the seminar by reading papers, running workshops, and participating in discussions. This gave language teaching practitioners an opportunity to listen and talk to many "big names" in testing. Several participants have indicated how interesting it was to listen to and meet persons whose work they have frequently read, cited, and applied.

The fourth opening was, no doubt, to some prominent aspects of the Finnish way of life, especially the sauna ritual. The male participants spent hours in the sauna, sweating and chatting, and on the command "lake, lake" all ran out and

plunged into the cooling waters of the lake, to return to the sauna for more of the heat treatment. This went on for hours. Like Coca-Cola, it was the real thing. We feel confident that the participants' cultural competence improved significantly and that the sauna ritual will be performed in a more criterion-referenced manner in the language testing community after the first-hand experience of participant observation.

The conference was also an opening for regularly recurring conferences and workshops in language testing to be arranged by the testing team in Jyväskylä. During the extensive self-assessment carried out by the University of Jyväskylä at the request of the Ministry of Education, followed by a week-long external assessment by a distinguished team of international experts, research on language learning and teaching was identified as one of the key areas in which good results have been obtained and which will be prioritized in resource allocation. Another area of strength was defined as evaluation and assessment of programmes, and language testing was specifically mentioned as an area of special competence.

Charles Alderson's review of the state of language testing provided a context for the conference and does so also for this publication. We have arranged the other contributions in three sections. The articles in Section 1 discuss themes related to language testing policy. Section 2 includes papers that have a mainly methodological thrust. Section 3 contains the majority of articles, which deal with a broad range of empirical topics on language testing.

We thank all the participants for their active contribution to the conference and especially those who revised their presentations to be included in this book. We are grateful to Mr. Jouni Hult for copy-editing the text for printing. We also wish to thank the Institute for Educational Research for publishing the book.

Jyväskylä, June 15, 1993

A.H., K.S., and S.T.

THE STATE OF LANGUAGE TESTING IN THE 1990s

J Charles Alderson
Lancaster University

The writer discusses issues in language testing that need attention in the 1990s. These include the use of qualitative methods of test analysis, the nature of judgements made by testers when constructing tests, new test methods, the content of language tests, and the impact of tests on language teaching and learning. Relevant literature and recent research regarding these aspects is reviewed.

1. INTRODUCTION

My title is impossibly ambitious: to attempt an assessment of the state of the art in language testing is doomed to failure, unless readers can accept that any such overview will inevitably be a personal statement, and perhaps reveal more about the author and that person's biases, interests and predilections than about the state of the art. In any case, other overviews exist which do the topic more justice than is possible or intended here: Skehan's review (Skehan, 1988) is an interesting and fairly comprehensive account of developments in the past decade. Volumes such as Anivan (1991) (in which Alderson, 1991a) and the papers in Alderson and North (1991) all address the broad topic, and it would be inappropriate to repeat the content of those papers here.

Instead, I want to concentrate on some of the issues highlighted in the above mentioned papers as needing attention in the 1990s. Other issues addressed in some of those papers — such as the nature of second language acquisition: what we know, what we don't know, how we try to find out and whether testing can help — are certainly important and will continue to be so. They will doubtless increasingly be addressed in the 1990s, but I do not propose to address them in this paper.

Nevertheless, before focussing on particular issues, it is worth reminding ourselves that developments in language testing have indeed occurred, and also that much remains to be done. Firstly, the amount of activity devoted to language testing has increased enormously in the past decade or so. Where once there were few publications and only occasional gatherings to discuss testing-related matters, now there are many publications, including journals and newsletters dedicated to testing, and international conferences on language

testing are very much on the increase — for some details, see Alderson (1991a). This increase in activity has doubtless increased not only knowledge about testing as well as an awareness of the limitations of our knowledge. It has also led to an increased professionalism in language testing: the recent foundation of ILTA — the International Language Testing Association — testifies to that. This increase in professionalism is certain to lead to an improvement in our professional standards, which ought to impact on the tests we produce and use (see Alderson and Buck, 1993 for a discussion of the need for this in the UK context). It is certainly to be hoped that this will, in the coming years, lead to an increased attention to the issue of standards for language tests and the development of commonly accepted procedures to reflect already accepted principles of language testing.

Secondly, it is my perception that there has been some shift in attitudes to language testing which has accompanied (and possibly related to) the increase in activity. Language testing used to be The Ugly Duckling of applied linguistics, much reviled, often ignored, but recently more applied linguists have taken an interest in the area and a recognition has grown that testing need not be divorced, either from teaching, or from applied linguistic theory (see Alderson and Clapham, 1991, for further discussion of the implications). As I shall discuss below, the development of language tests that attempt to reflect developments in language teaching, especially within 'communicative language teaching', has been a significant factor in this shift in attitude.

Yet it is also my perception that language testing has a long way to go before we can claim that attitudes to testing on the part of teachers and learners as well as of academics is uniformly positive. Tests are still seen by many as intrusive on the curriculum, as impacting negatively on learners, and, especially in the UK at present, as acting as an agent of control and Governmental policy, to the detriment of 'education' and educational 'values'. Whatever the truth of this latter belief, which is, like many policy matters, debatable, it is probably true that tests will always have negative connotations for some learners, and therefore teachers, as long as they have social consequences, i.e., as long as they are used to reward and punish, to select and reject. Although it is important constantly to remind ourselves that such uses of tests are beyond the responsibility of testers alone, we ignore test use and misuse at our peril. I shall be suggesting below that as language testers we have to date largely ignored the impact our tests might have, and this needs to be changed. If we pay more attention to conducting empirical research into the relation between teaching and testing in the next few years, we may well find that attitudes to language testing will also change, as it comes to be realised that language tests need not be scapegoats for the state of the art in language teaching.

2. ACHIEVEMENTS AND ATTENTION

Apart from an increase in activity and a perceptible shift in attitudes, how else has language testing developed over the past decade? In what follows, I propose to examine the state of language testing from the twin perspectives of what has been achieved, and what still needs attention. I shall address these two topics from a number of angles: analysis, method, content, and the relationship between testing and teaching.

2.1. Analysis

2.1.1. Statistical

Recent years have seen much interest in the development of new methods, particularly statistical, for test analysis. The advent of cheap but powerful microcomputers has been accompanied by the increased availability of software for computer-based analysis. Pre-eminent among these has been the family of methods belonging to what is known as IRT — Item Response Theory. Although IRT has been known for some time, and one variant of it, the Rasch model, was in use 20 years ago for the analysis of test results, it increased greatly in popularity in the 1980s, as its claimed advantages became well known.

Traditional test and item analysis has suffered from the fact that the estimation of a test's and an item's difficulty depended upon the ability level of the population of learners who responded to the test / item being analysed. However, IRT claims to be able to separate the difficulty of the item from the ability of the person taking that item, and therefore to be able to arrive at two separate measures: item difficulty and person ability. Theoretically, this makes possible the establishing of item and test difficulty independent of the trialling sample, and therefore it makes it much easier to develop tests of equivalent difficulty. The use of such forms of analysis has been the focus of many technical articles on language testing (see, for example, Language Testing, Vol 9, 2) and recent developments have included the extension of the partial credit model to enable the analysis of rating scales or non-dichotomous items.

In addition, multivariate statistics are increasingly used for test analysis, and causal modelling, including confirmatory factor analysis, has become relatively commonplace as a way of gaining a better understanding of what it is that tests are measuring. It is highly likely that we will see more developments of the use of sophisticated statistical tools for test analysis, and the continued exploration of their usefulness. However, the rate at which statistical tools and models have been developed has not, in my opinion, been matched by gains in insight as a result of the application of the tools. Methods like multitrait multimethod convergent- divergent validation have proved of some use, but lack rigorous

criteria that can be applied 'objectively' to determine validity. Confirmatory factor analysis has not replaced exploratory factor analysis as a tool for understanding relations among tests: indeed factor analysis itself is still generally considered to be of limited value in construct validation. The fact, of course, is that there is no 'alchemist's stone' of statistical analysis which can be used unfailingly and universally to establish test validity, whatever the adherents of a particular method at any point in time will have us believe. Fashions come and go in statistical analysis about as often as they do in clothing.

2.1.2. Qualitative

Statistical analysis has not been the only area where our analytical tools have developed. The 1980s was a decade in which interest developed, not only in applied linguistic research, but also in language testing research, in the use of qualitative techniques for gaining better insights into learners' test-taking behaviours. Cohen (1984) is an early example of using self-report techniques from learners to investigate test validity. From an initial interest in test-taking strategies, research has moved on to focussing on what learners' introspections can tell us about how the learner approaches a test item, what thought processes, strategies and test-wiseness techniques are engaged in by different learners, how do learners arrive at correct or incorrect answers? (Anderson, 1989, Anderson et al, 1991, Alderson 1990b, Buck 1990, Allan 1992, Li 1992).

What we are beginning to see, through studies like these, is that how individuals take test items varies considerably: the process is more complex than we often assume as test constructors and testing researchers. It is certainly more complex than is revealed by descriptive or inferential statistics based upon group rather than individual performances. The picture emerging is one of candidates using a variety of strategies/ techniques to arrive at their responses. The differences among these processes seems to depend upon variables which we can only guess at present, but they almost certainly include not only language proficiency and test-wiseness, but also personality, cognitive style, culture and background knowledge. These responses, the results of at times very different processes, are then judged as correct or incorrect by examiners, and inferences are drawn from scores to traits or abilities. However, the introspective results show very clearly that candidates can get items wrong, and yet have the abilities supposedly being tested by the item. They can also get the item right, and not display the required abilities. Indeed, the data show that test constructors are frequently wrong in their statements about what is being tested by test items. Tests evoke rather complex and variable behaviours and cognitive processes in individuals, which may be different for different individuals who make the same overt response. If the process is indeed variable, despite similarity of the product — response — then testers must be cautious in the claims we make for what it is that our items are testing.

This implies, at least, that fine-tuned diagnostic testing is going to be a difficult if not impossible task, since it means that there are a variety of appropriate ways in which somebody might have arrived at an answer. It also implies that the interpretations we make of test scores will have to be global rather than specific. That is, it might be legitimate to say that a candidate's score on a test of, say, reading tells us something about how well he or she has understood a text or even texts, and we may even be able to extrapolate from that score to an ability to understand other similar or dissimilar texts. But we may not be justified in making statements about the candidates' skills — enabling or micro skills to use Munby and Morrow's terms (Munby, 1978, Morrow, 1979) — that underlie that understanding. In other words we may be able to say something about the product of language use, but much less about the process, contrary to what many of us might want to be able to say or believed it was important to say. As our understanding of learner performance develops through the use of such qualitative research techniques, we come to see that it is highly variable, not only across individuals, but also within individuals across contexts, genres or settings (what Douglas and Selinker (1985) have called 'domains') We also come to see that the complexity of that variability makes it difficult and perhaps impossible to measure language processes by conventional test means. However, whether or not that is too pessimistic a statement, the point is that as testers we ignore information about process — as gathered by introspections as well as by other methodologies not yet explored — at our peril. The requirement of test validity demands that we investigate how candidates approach test items, however disturbing the results may be.

2.1.3. *Judgmental*

It is commonplace in language testing to make or gather judgements to help us decide on a test's content, method, standards, and validities. The commonest way of developing tests once items have been produced in draft form is to subject them to the scrutiny of 'experts': people, sitting on committees, inspect test items and make judgements about how well constructed they are, whether they match their specifications or the teaching syllabus, how difficult the items are, how appropriate for a given group of learners and, if unsuitable, how they might be modified in order to make them more suitable. In short, judgements are gathered about the items' content, difficulty and appropriacy. In addition, content validation is usually carried out by 'experts' making judgements about test content and its match with its purpose and its specifications. Tests are concurrently validated against judgements made by teachers as to learners' abilities and levels of proficiency. In other words, judgements abound in test construction and in test validation.

Sadly, however, recent research (Alderson, 1990a, Alderson and Lukmani, 1989, Buck, 1990, Allan, 1992) is beginning to cast doubt upon the accuracy of such judgements, and leads us to question a blind reliance upon judgements of content without further study. The evidence suggests that it is common for judges to disagree with each other and with test constructors about what a test

item is measuring. Some work, for example Bachman et al (1988), suggests that it is possible to improve judges' agreements about test content, partly by improving the framework or specifications that judges are asked to consider/ apply, and partly by careful training (see also Hughes et al, 1990). However, we need to be suspicious that such careful training is not simply cloning — getting people to think alike is not necessarily the same thing as proving that a test item really is measuring what the cloned judges agree it is measuring (see Alderson, 1991b for a further discussion of this issue).

Be that as it may, it does seem sensible for us at least to explore the agreement among judges as to test and item content, and to explore the reasons for disagreements. And since expert judgements form such an important part of the test construction process — especially at editing/ moderation stages — a study of the judgements made during the deliberations of such committees seems particularly important. In order to understand why our tests look the way they do, at least one thing we need to do is to investigate the nature of the editing and moderating process: the process of committee discussions and decision making, what knowledge bases and evidence are used in deciding on items, texts, item wording, test balance, and so on. How do such committees reach their decisions, what influences their judgements? This argues not only for examining the judgements people make, but also for looking at how and why they make those judgements, in order to see how 'valid' they are, and whether it is appropriate to devise ways through which better judgements can be made. It also argues for more ethnographic studies of what test constructors and committees do when producing a test (see Watson-Gegeo, 1988 and Hammersley, 1983 for excellent introductions to the use of ethnographic approaches in education).

One thing that is highly likely to influence decisions to delete, modify or retain draft items is judgements about item difficulty: it is my experience that test editors will frequently comment on an item's supposed difficulty and argue for and against item retention quite vigorously, believing in their own judgements of difficulty. Yet a study of people's judgements of difficulty compared with actual item difficulties (Alderson, 1993b) showed that judges are not able to predict item difficulty. I found that the accuracy of judgements was not associated with a judge's experience, either of test construction or of test marking, or even of the students taking the test. The evidence was that it was simply not possible to make reliable predictions of item difficulty. Yet it is at least conceivable that editing committees make decisions about test content on the basis of such predictions or intuitions. Interestingly, a study by de Jong (1984) has shown that the results of editing committees' adjustments to item difficulties after seeing item analyses of trialled items did not result in any significant improvement in item performance: as many items declined in performance on re-test as improved. Which suggests that experts sitting on committees may not be able to make accurate judgements at all. We need to look much more closely at the relationship between what committees and judges say about items and how those items actually perform. Conceivably,

feedback to committees about the performance of items might help educate and improve judgements.

2.2. Method

One of the major concerns in language testing for a long time has been with testing method: what methods can we devise to test different aspects of language use and language learning 'more' validly? What is 'The Best Method'? What do we think of multiple-choice as a test method? What does 'The Cloze Test' test? What are the latest methods? These and many more questions have dominated conferences, workshops and textbooks for a generation and more. Sadly for some but not surprisingly for others, no One Best Test Method has been found. Despite the claims made for the cloze test in the 1970s — which are still sporadically made in the 1990s — research has shown that it makes no sense at all to treat 'The Cloze Test' as a unified concept, other than trivially. It is not the case that a test produced by applying the cloze procedure will automatically result in a valid or even a reliable test (Alderson, 1979, Klein-Braley, 1981). Although the cloze procedure is convenient for test production — such that computers can easily be programmed to produce cloze tests — such convenience does not guarantee validity. Nor does the convenience absolve test constructors of the responsibility of deciding on the most appropriate method to test what they want to test, in relation to the learners being tested and their reasons for devising the test. Cloze tests need to be validated like any other test. Yet despite a general realisation among testing researchers that this is inevitably so, that any test produced by any method needs to be validated as a test for a given purpose, we have been diverted in the past decade by other Holy Grail methods: the c-test (Klein-Braley and Raatz, 1984) followed hot on the heels of the debunked cloze procedure, and was itself closely followed, at least for a brief period, by the 'cloze-elide' method (Manning, 1987).

Doubtless we will in the future continue to be treated to other messianic proclamations of the virtues of techniques not yet devised, or publicised. Test developers and test researchers alike must continue to be vigilant in their scrutiny of inflated claims for 'valid' testing methods. Research has clearly shown (Bachman and Palmer, 1982) that the test method makes a difference: i.e., that a learner's test score consists of both trait — that which we wish to measure — and method — the way we measure what we wish to measure, which is itself not what we want to measure. However, it does not follow that the influence of method will be constant across traits, item content, or indeed learners, and much more research will be needed into individual test methods before we can begin to have any confidence in predictions of 'method effect'.

One interesting development in the past few years has been a decline in popularity of the multiple-choice method. A popular test method because of the ease with which responses can be marked, multiple choice tests are very

common indeed in language testing. Yet it is not often realised that writing multiple choice tests is very far from easy: devising appropriate distractors which operate as the writer wishes them to is a skill (or art) learned only after long experience. much feedback from trialling items, and many mistakes. In addition, a growing dissatisfaction with the limitations of the method for testing language use rather than language knowledge has combined with insights from introspective studies and anecdotes of student test-taking behaviour and test-wiseness to result in a tendency among examining bodies to reduce their reliance upon multiple choice, and to explore other ways of testing language use, whether objectively or more subjectively. The short-answer question has increased in popularity recently, as have various gap-filling techniques (eg summary completion, as used in the IELTS test).

However, this decline in use of multiple choice has been paralleled by a revival of use of multiple choice techniques in computer-based testing (CBELT) as well as in computer assisted language learning (CALL). Clearly, multiple choice tests can be processed very easily by computer, but it is ironic that the advent of new technology has resulted in a revival of the use of a testing technique that has become less popular elsewhere. Indeed, since cloze tests are also easily produced and scored by computers, CBELT has also resulted in a revival of cloze-type tests. Alderson and Windeatt (1991) report on research conducted in order to investigate the potential of the computer medium for test methods that might be less traditional and less retrogressive. They show that even if tests have to be scored by computer, it is possible to develop a range of different item types whose use would at least avoid excessive dependence on out-moded item types like cloze and multiple choice.

Interestingly, Alderson and Windeatt's (1991) work has also resulted in a questioning of the distinction traditionally made between a test item and a learning exercise. It has been usual to believe that there is some sort of fundamental distinction between a test and an exercise, and the slogan has persisted in teaching circles for long enough that teachers should 'teach not test'. Yet Alderson and Windeatt argue that, once a tester can provide instant feedback on the correctness of a learner's response to an item, and can also provide hints and clues, test-specific dictionaries and grammars as further support, the distinction between a test and an exercise becomes exceedingly blurred. When the tester can also allow learners/ test-takers the opportunity to respond to the test item once more after receiving such clues and feedback, the distinction between a test and an exercise becomes possibly untenable. With paper and pencil tests, such feedback and subsequent 're-take' opportunities are practically impossible. When tests are delivered by computer through CBELT, this becomes perfectly feasible, and testers are then faced with the interesting dilemma: to provide feedback and subsequent re-take opportunities, or not? And if re-takes are allowed, should learners get full credit for a correct second response? Half-credit? quarter credit? How is one to make a principled decision? And indeed, which response is the most valid, providing the best picture of the learner's ability: the first or the second? Issues like these now

have to be faced when devising CBELT instruments, and the resulting discussion might well benefit the philosophy and practice of language testing in general.

However such dilemmas are resolved, it appears to Alderson and Windeatt (1991) that the only essential difference between a test and an exercise is that on an exercise a learner receives support, from a teacher or from peers, or from textbooks. On a test such support is traditionally termed 'cheating'. By giving learners the opportunity to 'cheat', is a CBELT test not a test? Yet if the provision of such support is the only difference between a test and an exercise, it would appear that language testers are free to be as creative in the item types they use as are exercise developers. This realisation can be genuinely liberating for item writers, who are all too used to thinking of writing test items as somehow different from what materials developers do. Yet there is no reason in principle why this should be, and it may well be the case that future years will see more imaginative use in language tests of item types that have been developed for use in 'communicative' language materials and classes.

One word of caution, however: item writing and materials writing may be very similar activities, but item writing is harder. Precisely because the learner does not receive support in a test, the instructions need to be absolutely explicit and clear whereas a teacher can rephrase, translate or demonstrate poorly worded exercise instructions. The teacher can skip ambiguous or poorly designed exercises, whereas learners are typically obliged to take all items in a test, and most items allow only a restricted number of correct or possible answers, whereas exercises can 'get away with' eliciting unexpected or irrelevant but correct responses because the teacher can, at least in theory, deal with the result. Exercise writing, in other words, is often sloppier than test writing is allowed to be (yet interestingly materials writers are not often criticised for such sloppiness, unlike test developers).

2.3. Content

The consideration of the relationship between testing methods and teaching methods, as related to computer-based testing and learning, brings us to the question of test content, since there is ample evidence that the content of language tests has been greatly influenced by developments in language teaching.

Since the 'communicative revolution', it has been commonplace to attempt to devise materials and activities which simulate 'real-life' language use activities. This desire for 'authenticity' in activities — in the texts to be read or listened to, to be spoken or written, as well in the tasks to be performed on such texts or in order to produce them — has had great impact on test content also.

Most language tests produced in the UK are believed, at least by their creators — see Alderson and Buck (1993) and Alderson, Krahnke and Stansfield (1987) — to reflect real-life language use. Texts are often facsimiles of real texts like TV schedules, tourist brochures, academic articles — and test takers are often expected to process such texts as they would if they were TV viewers, tourists or students. Test specifications and content are frequently drawn up with reference to a student's 'target language use situation'. In some cases this is indeed relatively predictable, and the 1980s saw the development of a number of ESP tests, particularly in the area of English for Academic Purposes, where test content attempted to reflect the student's likely future language needs. Just as many teaching materials were developed after some form of needs analysis, so too were some language tests, again most notably tests of English for Academic Purposes — see especially Weir (1983) and Carroll (1981).

Even for students whose linguistic needs were less predictable, emphasis has been placed in the past decade on devising tests whose content reflects plausible real-life language use — the former Royal Society of Arts Examination in the Communicative Use of English as a Foreign Language, and the Oxford Delegacy's Oxford Examination in English as a Foreign Language (Preliminary and Higher) are but two examples of this trend (see reviews in Alderson, Krahnke and Stansfield, 1987).

Unfortunately, the desire to replicate 'real-life' language use has led to 'authenticity' becoming something of a slogan in language testing as in language teaching. Whilst, other things being equal, it makes sense for test constructors to attempt to simulate real-life, there are many occasions when it is not necessary, or even sensible to do so. It is important to keep the purpose for which a test is being devised in mind, and it may be the case that the most appropriate way to achieve a given goal is not, per se, 'authentic'. Nevertheless, if a particular item elicits what one wishes to elicit and does so more effectively and efficiently than a more 'authentic' item, then it would be doctrinaire to go for the 'authentic' item rather than the inauthentic. For example, one may wish to test whether learners can exchange information orally using appropriate accurate English, using appropriate turn-taking strategies and paying due attention to, as well as understanding, their colleagues. One way to do this might be to develop a Spot the Difference task, where groups of four learners each have a picture, which contains two differences from everybody else's picture, and the task is for the group together to list the eight differences without showing each other their pictures. Most teachers would not consider such a task 'authentic'. Yet experience shows that it is capable of engaging students in information exchange much more effectively than a role play activity in which learners pretend to find out what time a train leaves for Bristol even though such an activity might appear more authentic.

The point I wish to make is that how well an item or a test works — measures what it is supposed to measure — is not always predictable. It is therefore essential to trial items in advance of their administration in the live test. If the

trial reveals that one item works better than another, even though it is inauthentic, then that item must be preferred to the authentic one. Similarly at the design stage, if it appears to test constructors that a relatively inauthentic task is more suitable than an authentic one, it would be foolish not to trial the inauthentic task. However, one would be well advised wherever possible also to trial the authentic task, since constructors can be wrong in their assumptions. Bachman (1990) distinguishes two types of authenticity: what he calls 'real-life' authenticity and 'interactional/ability' authenticity (IA authenticity). He says that the IA approach "focuses on what it sees as the distinguishing characteristic of communicative language use — the interaction between the language user, the context, and the discourse" (1990:302). He argues that the appearance of validity through 'real-life' authenticity should be subordinate to construct validity. The way that construct validity is achieved is through a theoretically adequate specification of what is to be tested, in other words, an adequate description of what it means to know a language. Whilst this is undoubtedly a tall order for anybody, even a language tester, it is clearly important that a language test be designed to measure language ability, and that implies that we can describe what language ability is.

Bachman has developed a model of Communicative Language Ability (CLA) from the original Canale and Swain model of communicative competence (Canale and Swain, 1980), to serve as a framework for test development. This CLA framework, which attempts to specify the components of communicative competence, is likely to influence the content of language tests for some time to come, at least until experience with operationalising the model reveals practical and theoretical problems that necessitate revision of the model. Already experience of applying the model to an examination of the Cambridge tests and the TOEFL has led to modifications to the model, but the importance of the model is that it exists, as a guide to test construction and to the operationalisation of applied linguistic theory. The development of the model is itself a significant contribution to language testing, and future attempts to apply and refine it will hopefully also be a contribution from language testing to applied linguistics more generally.

The past decade has seen a move in test content away from a focus on language structure and knowledge, away from a focus on accuracy through decontextualised, single sentence items, and towards attempts to measure learners' abilities to communicate. The focus of more recent 'communicative' tests has been on the message as much as the form, on measuring learners' abilities to perform in and with the language, on testing the ability to convey meaning appropriately and fluently. This has often been achieved through direct tests of performance, often task-based and reflecting real-life purposes of language use, informed by an analysis of what learners might have to do with language.

Nevertheless, there are many in language testing who believe that we must continue to measure a learner's command of the language, by which they mean

grammar. This does not amount to an advocacy of a return to discrete-point tests of knowledge, but it implies that attention needs to be paid to measuring whether learners can convey their meanings accurately. It is likely that language testers will need to look to developments in 'communicative grammar teaching' in order to develop suitable item types, and I would predict that we will see more focus in the future on defining and measuring 'grammar'. Given recent interest in understanding the nature of the lexicon, I would also predict that testers will need to pay more attention in the future to defining and measuring vocabulary.

However, another word of caution is appropriate: we need to beware of blindly following the fashions of language teaching, and the experience of the recent ELTS Revision Project is apposite, as recounted in Alderson (1993a). 'Communicative Grammar' and 'The Lexicon' are currently fashionable areas of interest; the ELTS Revision Project (the ELTS is a test of English for Academic Purposes — see the review in Alderson, Krahnke and Stansfield, 1987) was therefore charged with the responsibility of devising a 'communicative test of grammar', which would include tests of lexis, alongside tests of reading, writing, listening and speaking. We had considerable difficulty defining what was meant by communicative grammar, and how it might be tested, but we eventually produced a test that probably reflected the best of current teaching and thinking in this area: certainly it was the best we could produce in the circumstances. However, when we trialled this test alongside the other components of the new test, we discovered that it added little useful information to that provided by the other components. The correlations between the Grammar test and the other components were high, sometimes higher than the correlations between parallel tests of reading, for example. The reliability of the test battery was not affected by the removal of the Grammar test from the battery, yet by removing the Grammar test, we were able to save 30 minutes testing time in what was already a long test battery. We therefore dropped the Grammar test from the final IELTS battery. We concluded, in fact, that we were already testing 'Grammar' adequately through the other components: grammar is clearly needed in order to read academic texts, write, listen and perform well in an oral test!

3. TESTING AND TEACHING

In the previous section, I emphasised that recent developments in the content of language tests have been influenced by developments in language teaching. I want to take the discussion a little further by looking at other aspects of the influence of teaching on testing.

One significant development in modern language teaching in the UK has been the Graded Objectives Movement — GOML — which led to the development of Graded Tests. After the introduction of comprehensive education in the late

1960s, teachers of modern languages were faced by pupils with a much wider range of ability. Modern language teaching had previously been academic in orientation, grammatically based, and geared to traditional examinations after five years of instruction. However, this was irrelevant to many pupils in comprehensive schools, and the drop-out rate at the end of the two compulsory years of foreign language instruction alarmed many teachers. Not only that, but those pupils who did drop out had nothing to show for what little they had learned, since no tests were available. And so groups of teachers around the country decided to devise their own language tests at lower levels of achievement, which would be available to pupils long before the national examinations. These tests, which became known as graded tests, were often developed in a series of three or more levels, and pupils could progress through the series over a period of time. They were designed to test language relevant to the pupils' possible needs, as described above, but most importantly they were designed to be within the ability range of most if not all of the pupils learning foreign languages in comprehensive schools. Thus, they were designed to be achievable, in a relatively short space of time, and to provide pupils with something to show for what they had achieved. The result was a considerable success: drop-out rates were reduced, pupil motivation for foreign languages increased, and levels of achievement were generally thought to have improved. These graded tests clearly demonstrate the principle that learners can be motivated by having achieved something: success breeds success, whereas failure only breeds failure.

Graded tests also influenced the school leaving examinations in foreign languages, what became known as GCSE (the General Certificate of Secondary Education). Examinations were developed in the 1980s which were differentiated by level, to cater for a wider range of ability, which included an important element of teacher-based assessment — especially important for language tests which could then focus on the productive skills — and which were relevant to the needs of the pupils taking the exams. An important factor in the development of GCSE was in-service teacher education: teachers needed to learn how to conduct reliable and valid assessments, and this INSET doubtless had an important influence on how teachers taught as well as assessed. What was interesting about this development was the integration of teacher support with the needs of assessment, a topic to which we return below.

What these two developments illustrate is the importance of tests within education. Developments in the curriculum which ignore the tests and the assessment system more generally are likely to cause problems, developments which take the tests into account are more likely to succeed. Developing tests of an appropriate difficulty level, with appropriate content, delivered at an appropriate point within the curriculum, can have powerful effects. Indeed, this potential of tests for influence has long been recognised by educationalists. Unfortunately, it has also often be bemoaned: tests have been seen as having undesirable influence on the curriculum — the backwash or washback effect of tests on teaching. It is commonly asserted that tests have washback: that is, that

tests affect teachers and learners and thereby affect teaching and learning. (See, for example, Wiseman, 1961; Davies, 1968; Kellaghan et al, 1982; Alderson, 1986; Morrow, 1986; Pearson, 1988). Pearson (1988), for example, says:

> "It is generally accepted that public examinations influence the attitudes, behaviour, and motivation of teachers, learners and parents."

The belief in washback is most commonly asserted with respect to negative washback, namely the supposed undesirable effect on teaching/learning of a particular (and, by inference if not direct statement, "poor") test. In this case, 'poor' usually means 'something that the teacher or learner does not wish to teach or learn'. However, this may refer to the test's content — its syllabus — or to its method — the item types employed — or even to the way in which performances are scored/ marked.

It has often been observed that washback need not be negative: the term 'washback' implies influence, of any sort. If the test is 'poor', then the washback is felt to be negative. Logically, if the test is 'good', then its influence could be positive — if the Washback Hypothesis (WH) holds — namely, that tests influence teaching — then good tests should have good effects (as yet undefined). This has led some writers to advocate the use of tests to bring about change in teaching (see Pearson, 1988: 'Tests as Levers for Change').

Curiously, however, the belief in the power of tests to influence teaching has rarely been investigated: washback is frequently asserted, yet not studied. Alderson and Wall (1993) review the educational literature as well as writings in language education, and conclude that washback is seriously underexamined. They urge the research community to examine the nature of washback by observing what happens in classrooms when teachers prepare students for tests, and they argue that the Washback Hypothesis is over-simplistic, and is in need of considerable revision. They present a number of refinements to the basic hypothesis, which become progressively more complex as they take into account learning as well as teaching, the content and method of teaching/learning, the quantity and quality of teaching/ learning and so on. The fifteen hypotheses they advance are:

1) A test will influence teaching 2) A test will influence learning 3) A test will influence how teachers teach 4) A test will influence what teachers teach 5) A test will influence what learners learn 6) A test will influence how learners learn 7) A test will influence the rate and sequence of learning 8) A test will influence the rate and sequence of teaching 9) A test will influence the degree and depth of learning 10) A test will influence the degree and depth of teaching 11) A test will influence attitudes to the content, method etc of learning/ teaching. 12) Tests that have important consequences will have washback 13) Tests that do not have important consequences will have no washback 14) Tests will have washback on all learners and teachers 15) Tests will have washback effects for

some learners and some teachers, but not for others.

Wall and Alderson (1993) report on an attempt to do what Alderson and Wall (1993) advocate — a study of washback, in this case in Sri Lanka. Through classroom observation, as well as interviews with teachers, questionnaires, inspection of textbooks, Teachers' Guides and teacher-made tests, they conclude that the test they were investigating, a school-leaving GCE O-Level English examination, did indeed have enormous influence on teaching and teachers. However, they report that the influence was not as expected: the content of teaching was greatly influenced, but teaching methods were not. Furthermore, although teacher-made tests closely resembled the new exam, the way they were marked did not appear to. What Wall and Alderson (1993) show is that the phenomenon of washback is much more complicated than has been assumed. They argue that unless teachers are aware of the philosophy of an examination, and the principles on which it is built, they are unlikely to change their teaching methods. They conclude that for tests to have a positive effect on teaching, it is necessary to pay attention, not only to the design of the test itself, but also to the provision of teacher-support activities, including explicit advice to teachers on how to teach for the test. They also point out that 80% of the teachers who were interviewed in their study claimed that they were indeed influenced by the new exam. However, classroom observations showed that virtually no lessons showed any signs of impact on methodology. In other words, asking teachers if they are influenced by an exam will produce results that do not necessarily correspond to what actually happens in class. This argues for more research into washback, especially in classrooms.

Alderson and Wall (1993) suggest that what brings about change in classrooms — what makes teachers do what they do, or change what they do — is sure to be quite complex, and will relate both to personal factors such as motivation, as well as to institutional factors. They point out that the study of washback needs to be complemented by taking into account other factors that are known or thought to influence any innovation in teaching. They conclude that the Washback Hypothesis is unduly naive, and makes too many untested assumptions about how people are influenced. This applies as much to negative washback (eg the assertion that TOEFL forces people to do certain things) as it does to positive washback. The implication is that to blame tests for what happens in classrooms is quite inappropriate. Nevertheless, further research into washback is needed if the relationship between teaching and testing is to be explored and eventually better understood.

4. CONCLUSION

In this paper, I have not attempted to describe the major developments in language testing in the 1980s and 1990s. Rather I have given my personal view of what aspects of language testing have been important, and what is likely to remain important and to merit further attention in the future. In a sense, what I have produced is a research agenda for the next decade or so, which I very much hope will be as exciting as the last decade. I have tried to address a number of issues that bear on the validity, reliability and educational utility of the tests we produce, use and research. Areas that I believe will need attention in the coming years include:

- our methods of test analysis. They need to be expanded to include qualitative and ethnographic approaches to understanding what tests test and why tests are the way they are.

- the nature of the judgements we make as language testers. We need to investigate these more closely to see if they can be improved.

- the methods we use to measure student learning and linguistic proficiency. Reliance on a restricted number of methods is inappropriate, and we should be exploring the possibilities offered by new media and the insights obtainable from language teaching to develop our methods.

- the content of language tests. This should relate where appropriate to real-life language use but should not be dominated by an obsession with authenticity. Our choice of content needs to be informed by appropriate theoretical models of what it means to know a language, and this may well include redefining what we mean by 'knowing the linguistic system'.

- the impact of tests on teaching and of teaching on tests. There is clearly an important relationship between teaching and testing, but it is poorly understood at present and both would benefit from further investigation. In particular we need to research the impact of tests on teaching and learning, we need more systematic studies of what happens in classrooms, and how if necessary the educational impact of tests can be improved.

If we are able to achieve even part of this agenda, I believe that not only will a great deal more activity be generated, but also attitudes to testing will become more positive, and testing will receive more attention in language education and applied linguistics, to our mutual benefit.

REFERENCES

Alderson, J Charles. 1979. The Cloze Procedure and Proficiency in English as a Foreign Language. *TESOL Quarterly*, 13/2:219-227

Alderson, J Charles. 1986. Innovations in Language Testing? in Portal (ed)

Alderson, J Charles. 1990a. Testing Reading Comprehension Skills (Part One) *Reading in a Foreign Language* 6/2 pp 425-438

Alderson, J Charles. 1990b. Testing Reading Comprehension Skills (Part Two) *Reading in a Foreign Language* 7/1 465-5-3

Alderson, J Charles. 1991a. Language Testing in the 1990s: How Far Have We Come? How Much Further Have We To Go? in Anivan (ed)

Alderson, J Charles. 1991b. Letter to the Editor, *Reading in a Foreign Language*, 7/2

Alderson, J Charles. 1993a. The relationship between grammar and reading in an EAP test battery. In Chappelle C and Douglas D (eds)

Alderson, J Charles. 1993b. Judgements in Language Testing. in Chappelle C and Douglas D (eds)

Alderson, J Charles and Buck, G. 1993. Standards in Testing: A Study of the Practice of UK Examination Boards in EFL/ESL Testing. *Language Testing*, 10/1.

Alderson, J Charles and Clapham, C. 1991. Applied Linguistics and Language Testing: A Case Study of the ELTS Test. *Applied Linguistics*, 13/2

Alderson, J Charles and Hughes, A (eds). 1981. *Issues in Language Testing ELT Documents, 111.* London: The British Council

Alderson, J Charles, Krahnke, K and Stansfield, C (eds). 1987. *Reviews of English Language Proficiency Tests.* Washington, DC: TESOL Publications

Alderson, J Charles and Lukmani, Y. 1989. Cognition and Levels of Comprehension as Embodied in Test Questions. *Reading in a Foreign Language* 5/2, pp 253-270

Alderson, J Charles and North, B (eds). 1991. *Language Testing in the 1990s: The Communicative Legacy.* Developments in Language Teaching series, Modern English Publications.

Alderson, J Charles and Wall, D. 1993. Does Washback Exist? *Applied Linguistics*, 14/2

Alderson, J Charles and Windeatt, S. 1991. Computers and Innovation in Language Testing. In Alderson and North (eds) Allan, A (1992) *EFL Reading Test Validation: Investigating Aspects of Process Approaches.* Unpublished PhD dissertation, Lancaster University.

Anderson, N J. 1989. *Reading Comprehension Tests versus Academic Reading: What are Second Language Readers Doing?* Unpublished PhD dissertation, University of Texas

Anderson, N J, Bachman, L, Perkins, K and Cohen, A. 1991. An exploratory study into the construct validity of a reading comprehension test: triangulation of data sources. *Language Testing*, 8/1:41-66

Anivan, Sarinee (ed). 1991. *Current Developments in Language Testing* Singapore: RELC Anthology Series 25

Bachman, L F. 1990. *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press.

Bachman, L F and Palmer, A. 1982. The construct validation of some components of communicative proficiency. *TESOL Quarterly*, 16/4:449-465

Bachman, L F, Kunnan, A, Vanniarajan, S and Lynch, B. 1988. Task and ability analysis as a basis for examining content and construct comparability in two EFL proficiency tests. *Language Testing*, 5/2:128-159

Brumfit, C J and Johnson, K (eds). 1979. *The Communicative Approach to Language Teaching*. Oxford: Oxford University Press.

Buck, G. 1990. *The Testing of Second Language Listening Comprehension*. Unpublished PhD dissertation, Lancaster University

Canale, M and Swain, M. 1980. Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, 1/1:1-47

Carroll, B J. 1981. Specifications for an English Language Testing Service. in Alderson and Hughes (eds)

Chamberlain, D and Baumgardner, R (eds). 1988. *ESP in the Classroom: Practice and Evaluation*. ELT Document 128. Modern English Publications

Chappelle, C and Douglas, D. (eds). 1993. *A New Decade of Language Testing Research*. Washington, DC: TESOL Publications

Cohen, A. 1984. On Taking Language Tests: What the Students Report. *Language Testing*, 1/1:70-81

Davies, A. 1968. *Language Testing Symposium: A Psycholinguistic Approach*. Oxford University Press

Douglas, D and Selinker, L. 1985. Principles for language tests within the 'discourse domains' theory of interlanguage: research, test construction and interpretation. *Language Testing*, 2/2:205-226

Hammersley, M (ed). 1983. *The Ethnography of Schooling*. Driffield: Nafferton

Hughes, A, Porter, D and Weir, C J. 1990. *Reading in a Foreign Language* 7/1

de Jong, J H A L. 1984. *Het rendement van pretesten*. Interne Documentatie Nr 126. Arnhem: CITO

Kellaghan, T, Maduas, G F and Airasia, P W. 1982. *The Effects of Standardized Testing*. London: Kluwen, Nijholf Publishing

Klein-Braley, C. 1981. *Empirical Investigations of Cloze Tests: An Examination of the Validity of Cloze tests as Tests of General Language Proficiency in English for German University Students*. Unpublished PhD dissertation, University of Duisburg

Klein-Braley, C and Raatz, U. 1984. A survey of research on the c-test. *Language Testing*, 1/2:134-146

Li, Wen Yao. 1992. *What is a Test Testing? An Investigation of the Agreement between Students' Test-Taking Processes and Test Constructors' Presumptions*. Unpublished MA thesis, Lancaster University

Manning, W H. 1987. *Development of Cloze-Elide Tests of English as a Second Language*. Report 23, TOEFL Research Reports series. Princeton, NJ: ETS

Morrow, K. 1979. Communicative language testing: revolution or evolution? h Brumfit and Johnson (eds)

Morrow, K. 1986. *The Evaluation of Tests of Communicative Performance in Portal (ed)*

Munby, J. 1978. *Communicative Syllabus Design*. Cambridge: CUP.

Pearson, I. 1988. Tests as Levers for Change. in Chamberlain and Baumgardner (eds) pp98 - 107

Portal, M (ed). 1986. *Innovations in Language Testing*. NFER/Nelson.

Skehan, P. 1988. Language Testing, Part 1 and Part 2: Survey Article, *Language Teaching*, p222 - 221 and 1 - 13

Wall, D and Alderson, J Charles. 1993. Examining Washback: The Sri Lankan Impact Study. *Language Testing*, 10/2

Watson-Gegeo, K A. 1988. Ethnography in ESL — defining the essentials. *TESOL Quarterly*, 22/4 pp 575-592

Weir, C J 1983. *Identifying the Language Problems of Overseas Students In Tertiary Education in the UK*. Unpublished PhD dissertation. University of London Institute of Education.

Wiseman, S (ed). 1961. *Examinations and English Education*. Manchester University Press.

Section I

Language Testing Policy

THE EXERCISE OF POWER AND CONTROL IN THE RHETORICS OF TESTING

Elana Shohamy
Tel Aviv University

The article argues that tests are used to exert power and control over test takers. The writer brings forth evidence from the rhetoric and discourse used in testing to back up the argument. Further support is provided by the gap that appears to exist between what is said about tests and how tests are actually used.

INTRODUCTION

Tests are used for a variety of purposes — to measure students' knowledge in relation to future tasks which they are expected to perform, to place students in appropriate levels of classes, to grant certificates, to determine whether students can continue in future studies, to select those who are most suitable for higher education institutions, to obtain jobs and to accept candidates to different programs. Thus, results obtained from tests have a strong impact on the lives of individuals. The Irish novelist McGahern (quoted in Madaus, 1991) describes the impact of tests, taken to an extreme, of a child who is about to take a test, in the following poem:

> Please God may I not fail
> Please God may I get over sixty per cent
> Please God may I get a high place
> Please God may all those likely to beat me get killed in road accidents and may they die roaring.

This strong impact of tests on test takers has turned tests into powerful and authoritative devices which are capable of changing and prescribing the behavior of those who are affected by their results. A phenomenon that can be observed as a result, the exercise of power and control through tests can be observed in a variety of settings and contexts. Tests are used by policy makers in central agencies as tools to manipulate the educational system, to control curricula and to impose the introduction of new textbooks and new teaching methods. Principals use tests to prompt teachers to teach, and teachers use tests to force students to study. On the national level power is exercised by bureaucrats through nation-wide tests such as the matriculation examinations and national assessment surveys; on the school level control is exercised by

principals and administrators through school-wide exams; and in the classroom power is exercised by teachers through tests, quizzes and grades. The power and authority of tests turns them into tools for enforcing educational goals especially in situations when actual education has failed. This phenomenon can be observed in all levels of education — national, school-wide and classroom.

Madaus (1991) states that tests represent a social technology deeply embedded in education, government and business and as such provide the mechanism for enforcing power and control. Foucault (1979) in *Discipline and Punish: the Birth of the Prison*, views the examination as the most efficient tool through which society imposes discipline as it contains all the features needed for power and control.

> The examination combines the technique of an **observing** hierarchy and those of **normalizing judgement**. It is a **normalizing** gaze, a **surveillance** that makes it possible to **quantify**, to **classify** and to **punish.** It establishes over individuals a **visibility** through which one **differentiates** them and judges them. That is why, in all the mechanisms of discipline, the examination is highly **ritualized**. In it are combined the **ceremony of power** and the **form of the experiment, the deployment of force** and the **establishment of truth.** At the heart of the procedures of disciplines, it manifests the **subjection** of those who are perceived as **objects** and the **objectification** of those who are **subjected** (1979:184).

This paper will bring evidence from testing discourse to show how tests are used for power and control, specifically for observation, surveillance, quantification, classification, normalization, judgement, and punishment. Evidence will be reported from two main sources: a) rhetorical and discourse devices used in tests and testing, and b) the gap existing between the discourse and intentions of decision makers regarding tests and the specific actions observed in actual test use.

1. RHETORICAL DEVICES AS INDICATION OF POWER AND CONTROL

1.1. The Channel of Communication

The first rhetorical device indicative of the power and control of tests relates to the channel of communication. Most tests used today are written examinations. Madaus (1991) claims that it was not until the latter part of the 18th century when paper was more readily available that written tests were systematically introduced into European schools, replacing the traditional oral mode of testing. Foucault (1977) contends that **written** examinations mark the beginning of scientific pedagogy. The written test defines what is expected and forces

students to reveal periodically how their learning is progressing. It "guaranteed the movement of knowledge from the teacher to the pupil, but it extracted from the pupil a knowledge destined and reserved for the teacher, thus becoming a mechanism for exercising power over the pupil". (1979, p. 187).

Earlier tests which employed the **oral** mode, left room for a dialogue between a tester and a test taker — the student had the opportunity to ask questions, to elaborate on what s/he had learned and to demand clarification. The written test introduced a form of discourse which altered the act of communication from a dialogue to a one way channel of communication focusing on what the teacher had taught without considering the test taker's views and interpretations. Thus, while the oral mode is characterized by negotiation, elaboration, expansion of ideas, use of contextual clues, correction, argument, revision, debate, the reading of facial and contextual clues, etc., the written mode is one-sided; information flows in one direction as the test takers writes what he knows in response to his interpretation of a written stimulus.

1.2. The Power of the Language of Numbers

Another rhetorical device which contributes to the power and control of tests is the language of numbers through which test results are disseminated. Numbers are powerful since they allow quantification, classification, normalization and judgement. Much has been written about the use of numbers and statistics as effective devices of power and control and about the fact that they speak more than words (McLuhan, 1983). Bureaucrats, principals and teachers use numbers to classify, to punish, to reward, to place, to accept and to reject.

Those who work with statistics and numbers know how unreliable numbers are, but those affected by numbers have trust in their power and authority. The power of numbers also lies in the fact that they can be challenged only if one has numbers which counter them. In testing, however, the tester has the numbers while the test taker does not, and the only numbers he can refer to are those provided by the tester who "owns" the numbers.

The power of numbers also lies in the impression that they are objective and therefore represent the "truth"; they are not open to discussion and challenge. Numbers are powerful also because they allow classification and standardization on the same scale and thus enable all test takers to be judged according to the same yardstick. Thus, the language of numbers as a device for disseminating test results provides decision makers with power, authority, control and the legitimacy for decisions and sanctions.

1.3. The Power of the Language of Science

The language of science is viewed as a tool for power and control. The fact that testing is perceived to be a scientific and empirical discipline grants testing authority and status. Testing is powerful because it is believed to be scientific, experimental, statistical and therefore objective and true.

The "objective" and "empirical" information gleaned from tests is then used by bureaucrats to support their beliefs. MacIntyre (1984) states that since the aim of the bureaucrat is to adjust means to ends in the most economical and efficient way he will deploy scientific knowledge organized in terms of and comprising a set of universal law-like generalizations to help him. Schwandt (1989) elaborates on this phenomenon and claims that bureaucratic managerial expertise in both the public and the private sectors rests on this assumption of scientifically based practice. Evaluators, policy analysts, researchers and consultants enhance bureaucratic expertise by acting as scientific resources.

> As masters of the tools of scientific analysis, these experts generate empirical evidence — scientifically based knowledge — that managers use to enhance their authority and control. (Schwandt, 1989:12).

Although testers know that they are unable to generate accurate generalizations concerning human behavior in view of the constraints imposed by real-world problems on methods of scientific investigation, the belief in the role of the evaluator as an expert who provides scientifically derived knowledge for managers and policy makers remains essentially intact.

1.4. The Power of Written Documentation

Results obtained from tests are documented. Foucault (1979) writes that the examination which places individuals in an area of surveillance also situates them in a network of writing; it generates a whole mass of documents to capture and fix them. Thus, the power of writing constitutes an essential part in the mechanism of disciplines. Specifically he states that not only is the aptitude of each individual assessed, but the procedure of writing also situates these abilities and indicates the possible use that might be made of them.

> Thus, the procedure of writing made it possible to integrate individual data into cumulative systems in such a way that they were not lost, so to arrange things that an individual could be located in the general register and that, conversely, each datum of the individual examination might affect overall calculations. Thanks to the whole apparatus of writing that accompanied it, the examination opened up two correlative possibilities, firstly the constitution of the individual as a describable, analyzable object,

> not in order to reduce him to "specific" features, as did the naturalists in relation to living beings, but in order to maintain him in his individual features, in it his particular ovulation, in his own aptitude or abilities, under the gaze of a permanent corpus of knowledge, and secondly, the constitution of a comparative system that made possible the measurement of overall phenomenon, the description of groups, the characterization of collective facts, the calculation of the gaps between individuals, their distribution in a given "population". (Foucault, 1979:191)

Combining the tests with documentation implies that what happens on a test is recorded, will remain and will provide those who are interested in the information with evidence for possible sanctions against test takers. Thus, the documentation allows the transfer of the procedure of writing and registration into the formation of a mechanism for discipline.

1.5. The Power of the Objective Test Format

Test format relates to the genre in which test questions are posed. Most tests used by testing agencies employ "objective" items such as multiple choice and true false. The assumption underlying these procedures (and the reason they are called "objective") is that there is one correct answer, i.e. one "truth", one interpretation. Furthermore, that "truth" is owned by the tester and the test taker has to try to find out what it is. Objective testing controls knowledge by creating one "truth" with no deviation and no alternatives; it provides power and control through the format of the test; the "truth" is absolute and determined in advance by the tester who wrote the test and selected all the alternatives. It does not ask the test taker to present his interpretation or any other possible interpretation of the truth. This is especially problematic in the testing of reading comprehension where there may be many interpretations of a similar text, as readers construct meaning differently depending on their background, age, views, etc.

Thus, an "objective" format means control since the information is in the hands of the tester who writes the item, selects the correct answer according to his/her interpretation and allows no deviation from it. Since comprehension is subjective, the "objective" format of testing demands a type of knowledge which is an inaccurate reflection of "truth".

The next part of the paper will describe the gap that exists between how decision makers "talk" about tests and the actual use made of tests for power and control.

2. THE GAP BETWEEN WORDS AND ACTION

The comparison between the stated intention of the tests and their actual use provides evidence to support the claim that tests are used for power and control. Tests are described by decision makers as useful, educational, beneficial and significant for the advancement and improvement of education; yet in examining their actual use there is ample evidence to show that what is said is not what is done, and what is said often hides a different message, that of power and control. Specifically tests are used for observation, surveillance, quantification, classification, normalization, judgement, and punishment. A specific case of a national test administered on May 27th, 1991 in reading comprehension and math for fifth and sixth grades in one country will be used to exemplify this point. This test is typical of other national tests given in many countries for monitoring achievement, especially in countries which have a centralized educational system, e.g. a national curriculum and national exams for different age groups. This specific national test was administered to 160,000 students, in 6000 classes, in 1600 schools, and employed 6000 external proctors. Prior to the administration of the test the Ministry of Education sent the following letter to the parents of the students who were to be tested:

> Dear Parent:
>
> The Ministry of Education and Culture is about to administer tests in Math and Reading Comprehension to all the students. The results of the test will enable us to find out about the level of achievement of children in these important subjects. The tests will help the school plan its work for the next academic year. The results will be used by the Ministry only for pedagogical purposes, for tracking, for research and for establishing policy. There will not be any use of the results for purpose of selection. The data will be kept in the data base of the Ministry of Education in accordance with the Privacy law (1981). The detailed results will be sent to the school after the beginning of the next academic year.

What can be more educational than that? In the next sections the consequences of the tests in terms of power and control will be described.

2.1. Using Tests for Surveillance

The first indication of the use of tests for power is using them for surveillance. The reading comprehension test was admittedly an act of surveillance by the government. The Ministry of Education follows the progress of all the children of the country. It can, without asking permission of parents, students or teachers, compel them to be tested; it can enter the classrooms and demand that

all the children in the country take a test and show the Ministry what they know in a certain subject according to the Ministry's criteria. If in fact the Ministry did not intend to use the results of the test for the selection of individuals, why were students asked to write their names? why were the scores released to schools by students' names? In administering the tests the Ministry did not have to prove that the tests are not "harmful to the students or that these tests are of such high quality that they provide accurate scores. Students or parents have no right, and in most cases no tools, to examine the quality of the test before-hand. All students must participate and if some students refuse, they or their teacher may be punished. The performance of each of these students is then recorded in the form of a number which is classified as pass or fail. The government now has records on each of its pupils (including thousands of newly arrived immigrants for whom the language of the country is a newly acquired language).

The information that the Ministry can obtain from this surveillance cannot be of very high quality since the decision on who failed and who passed was based on a mere 15 test questions. The results do not take into consideration the relevance of the texts to the background of the students, their social and ethnic environment, their home culture, and especially the environment of the immigrant population, nor whether the student had a good or bad day when doing the test.

It is interesting to note that in this specific case, not only has the government the right to force all student to be tested, it can, and in fact, did hire a squad of proctors to help in the surveillance, that is, it actually demanded that students from the different teacher colleges around the country become proctors to make sure that the students not cheat, while the classroom teachers were forced to leave the classrooms during the test. The government and the Ministry are powerful enough to demand that students from different institutions miss a day of studies in order to participate in this surveillance mission.

But not only are students surveyed, so too are teachers and schools. The documented, recorded and classified information relates to teachers as well, as if it depended only upon the teachers whether their students succeeded in or failed the reading comprehension test.

2.2. Using Tests for Quantification

Students and teachers are now recorded in the data base of the Ministry of Education in the form of quantification, i.e. their identity is now described, defined and summarized in terms of "percentage of correct answers" on a 15-text test.

2.3. Using Tests to Classify People

In addition to the ability to test anybody, anytime, and to quantify them another aspect of power and control is the classification of students as successes or failures, as good or bad. Thus, it is possible for the authoritative group to classify students, teachers and schools into categories. Often, the students, schools and teachers are not even aware that they have been classified as successes or failures.

Evidence for the use of tests for classification, tagging and punishment can be found in many tests administered on a regular basis. The first time the Ministry administered a nation-wide test in Arabic, students were asked not to write their own names but only the name of their teachers.

2.4. Using Tests for Judgements and Sanctions (Punishment)

Once the information is available (observed, documented, quantified and classified) there is legitimacy for imposing sanctions. Examples of sanctions are forbidding students from entering certain class levels, demanding that they repeat grades and withdrawing from upgrading programs.

Even if the Ministry claims not to use this information for selection, the availability of the information implies that it will be used. Schwandt (1989) claims that bureaucrats are looking for such information in order to justify their actions and will therefore make use of it. There are many examples of situations were teachers arrive at decisions based on test results which were meant for entirely different purposes. In some schools, teachers were fired, textbooks were changed and students were placed in low level groups based on tests which were initially intended only to evaluate the level of the schools (Shohamy, 1991).

The difference between description and judgement is relevant in this context; while tests give descriptive information, decision makers often use the results for judgement, i.e. punishment or reward. It is hard to believe that a single test score can result in major decisions and consequences for people's lives. Madaus (1991) claims that rarely do we find a single device that is so powerful that it can dictate so many decisions as is a single standardized test score which independently triggers automatic admission, promotion, placement or graduation decisions. These decisions are non-negotiable even in the face of contradictory judgements from educators about what a student knows or can do. Such reliance on a single test score is a process vulnerable to simple error. In the area of policy, a test is a social technique which yields information that can be used to direct people in different directions. Thus, using a single test score can put children at risk.

2.5. Using Tests for Scaling and Standardizing the Entire Population

While each of the students tested on the RC test was an individual case before the test, after the test s/he was judged in relation to all the other students of his/her age in the country. It is important to mention that the norm for the country is determined, in most cases, by how the rest of the country does on the very same reading comprehension questions, as the same test is given throughout the country. Thus the test has the power to straighten the level of deviation in the country. Whereas this year there may have been some deviations as a result of different backgrounds, different teachers, different learning contexts and different ways of constructing meaning, the effect of this test will ensure that this will not happen again and that students around the country will construct meaning in the same way and will end up with exactly the same knowledge. In countries which have centralized educational systems this is a well known phenomenon since any national test ensures that all students will arrive at the same knowledge and the test plays a major role in this standardizing and norming effort. This phenomenon of standardizing knowledge throughout the country based on the same test questions is exemplified in Figure 1, taken from Foucault's book *Discipline and Punishment* (1979).

Figure 1. **N.Andry, *L'orthopèdie ou l'art de prévenir et de corriger dans les enfants les difformités du corps* (Orthopaedics or the art of preventing and correcting deformities of the body in children). 1740.**

Though the above phenomenon has long been typical of countries with centralized educational systems although in the past few years there has been a similar trend even in countries with decentralized educational systems. In the US, for example, the ACTFL guidelines, often referred to as a "common yardstick" were introduced to define the core of knowledge to be acquired and tested by all foreign language learners as part of the effort to upgrade the proficiency level in foreign language learning. These guidelines imply one type of foreign language knowledge; deviations from it need to be corrected. This is an act of control and power as people are being controlled and standardized according to a single scale. While it is clear from language learning theories that there are many routes to acquiring a foreign language; the common yardstick concept implies standardization to one single path. It is interesting to note that when such an act occurs in countries with decentralized systems based on freedom of education, the government uses scales to circumvent that underlying assumption.

2.6. Using Tests to Demonstrate Authority (who is in charge)

Tests are often used for the purpose of declaring where authority lies; this is especially true in centralized systems. In many developed countries responsibility and authority are shared by one body. In some countries there is a clear distinction between authority and responsibility.

The tension between the authority and those who are responsible for introducing the change is clear in the case of the national reading comprehension test. One wonders whether teachers (those responsible for change) will be willing to cooperate with the Ministry of Education (those in authority) in the future. The teachers were humiliated by the system which viewed them as potential cheaters and not trust-worthy when they were forced out of their own classrooms during the test and not briefed on what their students were expected to know. One also wonders what kind of message is delivered to students about their teachers. It is after all, the teachers who will be asked to cooperate in "correcting" the various reading comprehension problems detected on the tests.

Another strategy to demonstrate who is in control is the "defense mark" used in a number of countries as part of the matriculation examinations administered at the end of secondary school. The defense mark, assigned by the students' teacher, is one component of the matriculation score. The students' final score consists of the average of the external matriculation examination and the "defense mark." The term "defense mark" already implies that the student needs to be "protected" from the system, from the authorities. But if the school deviates from the external score, it receives a warning from the authorities based on the assumption that the external score (of a once in a life-time test) is "the correct score" and that the Ministry of Education which administers the external test provides the most valid assessment while all other grades,

especially those assigned by the teacher (who may have known the student for a number of years and in a variety of contexts) are not valid and the teacher must conform to the external scale. Thus, if gaps occurs between the two scores, the school receives a warning that the school(!!!) does not assign valid grades as the Ministry owns the "truth".

It is often the case that tests are conducted by Ministries of Education when they feel that their authority has been threatened, or when they feel that they are losing control; the use of the test helps bring the authority back to power. The administration of national tests of any kind is often aimed at indicating who is in charge. This is especially true when the system fails on other fronts. The reading comprehension test discussed above was administered by the Ministry of Education to a large extent as a reaction to complaints that innovations and new programs in reading comprehension were not being introduced into the system. Similarly, the introduction of the ACTFL guidelines in the US was a reaction to criticism that there were no new programs in foreign language teaching and that children were not performing as expected. Though no new programs have been introduced recently, the administration of the test is an indication to the public that some things are being done.

2.7. Using Tests to Frighten and Deter

Using tests to frighten or deter is related to the use of tests by the authoritative agencies to stimulate the system. It usually occurs when there is failure in learning. Thus the current proposals to introduce national tests in the USA, are a result of the low scores that American children obtain on international tests (Madaus, 1991). There are a number of cases where the Ministry of Education introduced tests in order to "shake up" the system. It is interesting to note that this goal of shaking the system is not done through the introduction of new curriculum but through tests. This is a result of the power of tests and the fear that students and teachers of have of failing.

There are various techniques the Ministry employs to enhance that fear and make it more effective. Before the national reading comprehension test, the Ministry of Education used the strategy of not making information available to teachers and students; it refused to release any information about the test, the format, the questions and its purpose. Withholding the information, not unexpectedly, created tension, anger and anxiety. One reason that teachers were asked to leave the classrooms was not to reveal the questions so that the same test could be used the following year. However, copies of the test were transmitted all over the country. The strategy of keeping information from the public is often used by teachers who try to avoid releasing information about tests to students and thus to exercise power and control. The element of mystery is strongly enhanced and is related to the ritual phenomenon mentioned by Faucoult (1979) which is typical of the use of disciplinary tools.

Absence of information about the tests is a device testers use to impose power and control. By withholding information from the public at large and the test takers in particular, the situation created is that of the tester owning the knowledge which the test taker lacks. That knowledge gives the tester extreme power and leaves the test taker with no power.

However, newspaper reports indicated that in spite of the Ministry of Education's intention, students spent months preparing for the test; poor students were removed from classes in fear that they would lower the school average, good students were not excused from school even if they had a previous important engagement.

2.8. Using Tests to Control Learning

Testers claim that the nation-wide test assesses what the students have learned. However, it turns out that the testers dictate **what** will be tested and consequently what will be learned. Shepard (1991) claims that testers view tests as synonymous to curriculum and to learning. Nation wide tests are often based on material that not **all** schools were exposed to.

Thus, a decision to include a certain language skill (e.g., speaking) or alternatively, not to include another area (e.g., grammar) on a high school matriculation exam, will result in students and teachers practicing speaking and not being motivated to practice grammar. Thus, tests control the knowledge of students by determining that they learn the topics considered important by testers.

Consider the example of a national supervisor of language who felt that students graduating from high school lack proficiency in reading comprehension and writing. In order to 'cure' this problem he decided to change the national exam at the end of high school by adding reading comprehension and writing components. As a result, teachers demanded more hours to teach the content, new textbooks were commercially produced, children were tutored in these skills, and a major effort was put into the study of an areas which had not been taught before, all in preparation for the important test. Interestingly, reading and writing have always been included on the national curriculum, but since they had never been **tested** on a national exam they apparently had not been taken seriously by teachers and students.

3. IS THE USE OF TESTS FOR POWER AND CONTROL EFFECTIVE?

Since the use of tests for the purpose of power and control has become such a wide-spread phenomenon in many countries, questions have been raised not only regarding the ethics of using tests for these purposes but also whether

using tests in such a way can advance and improve learning in a meaningful way. A number of studies have shown that prescribing behavior is not an effective way for upgrading the level of learning (Stake, in press; Smith, 1991). This leads to skepticism about the extent to which the introduction of tests, per se, can bring about meaningful improvement in learning and teaching. Consequently, there are serious doubts whether it is at all beneficial to rely on the power and authority of tests to create change in the behavior of teachers and students.

Stake (in press) and others studied the effect of introducing tests as devices for upgrading achievement and found that although the tests did make teachers more focused in their teaching, they resulted in **narrowing** the scope of the subject matter because teachers were teaching primarily for the test.

Similarly, Shohamy (1989) examined the effect of the introduction of a new EFL (English as a Foreign Language) oral test as part of the nation wide high school matriculation exam, on the level of English proficiency and found the impact to be that of teaching "test language", that is, only those language tasks and skills likely to appear on the tests were taught in English classes. Furthermore, teachers tended to shift their teaching from the means and procedures that **facilitated** learning, to the goals. Some teachers admitted returning to "normal" teaching right after the administration of the test; they clearly viewed the test as an external, irrelevant act in relation to their routine teaching.

Such findings should not come as a surprise. After all when there is exclusive reliance on tests to create change; when the emphasis is not on what takes place in the classroom as part of the learning process; when tests are introduced as authoritative tools, are judgmental, prescriptive, and dictated from above; when the tests do not involve those who are expected to carry out the change — the teachers; and when the information provided by tests is not detailed and specific and does not contain meaningful feedback and diagnosis that can be used for repair, it is difficult to expect tests to lead to meaningful improvement in learning. Tests, after all, comprise only one component in the educational process.

4. INCORPORATING TESTS INTO THE EDUCATIONAL SYSTEM

Creating change through tests can, however, be effective, since schools will usually strive to meet external standards and will alter teaching and learning habits to attempt to improve performance on tests. At present, however, curricular change does not result; only the performance of the students on these tests is emphasized. The effect of this is to narrow the curricula in ways which are inconsistent with real learning and real needs.

The main question, is therefore, how the device of testing can be utilized so as to create positive effects on learning. One solution offered today is to maximize the role of tests as a means for obtaining information leading to th improvement of learning while minimizing their role in power and control. The real power of tests lies not in their authority and prestige, but rather in their ability to provide meaningful and relevant information which can be incorporated into the educational system. Thus, there is a need to rely less on the power of tests and more on the useful information that tests can provide for the improvement of teaching and learning. It is possible, for example, to use tests to supply decision makers (students, teachers, and administrators) with valuable information and insights on teaching and learning. The information obtained from tests, when it is relevant and innovative, can provide evidence of teachers' performance and of students' ability over a whole range of skills, subskills and dimensions, achievement and proficiency, on a continuous basis and in a detailed and diagnostic manner. According to Nitko (1989) information that is innovative and meaningful can lead to utilization. Such information can be useful for judging students' in relation to the expectations outlined in the curriculum, to determine whether the school as a whole is performing well in relation to other schools that share the same curricula, to determine whether the teaching methods and textbooks used are effective tools for achieving those goals, and to determine whether the goals are realistic and appropriate. It is the **information** obtained from tests and its utilization which should make them valuable and powerful. When decision makers realize the value of the information tests provide improvement can result. Information can be utilized properly especially when teachers, those expected to carry out the changes, are involved in the assessment process. Once conclusions are reached based on this information, it is possible to align the curriculum with the information and to implement changes in goals, teaching methods and textbooks. These changes can then be monitored through repeated administration of tests on an on-going basis.

Fredriksen and Collins (1989) contend that information obtained from tests can contribute to improved learning when tests are connected to the learning that takes place in the classroom. They introduce the notion of **systemic validity** to refer to the introduction of tests together with a whole set of variables which are part of the learning and instructional system. Accordingly, the introduction of tests is a dynamic process in which changes in the educational system take place according to feedback obtained from tests. In systemic models, a valid test is one that brings about, or induces, an improvement in the tested skills after a test has been in the educational system for a period of time. Fredriksen and Colling contend that high systemic validity can only be achieved when a whole set of assessment activities foster it. They identify such activities as the use of direct tests, practicing self assessment procedures, repeated testing, feedback on test performance and multiple levels of success. They claim that the efficiency of current testing practices is greatly outweighed by the cost of using a system that has low systemic validity — one that has a negative impact on learning and teaching since the goal of testing has to be to support the improvement of

learning and teaching. Thus, tests that are used for the purpose of improving learning can be effective when they are connected to and integrated into other elements which are part of the educational system and not when used in isolation.

Testers must begin to examine the consequences of the tests they develop. Testers often feel they have completed their job after obtaining a high reliability and validity and do not find it necessary to observe the actual use of the test. It has been claimed (Messick, 1989) that it is the responsibility and obligation of the testers to examine the uses, misuses and abuses of tests in the real world. Their role is to warrant against such misuses described in this paper and to show that in spite of the educational purposes of tests, they may actually be used for power and control.

It is essential that testers cooperate with experts not involved in testing and acquaint all those concerned with tests and their use with the vocabulary of testing so that they can enter the discourse without being dismissed as naive, thus improving communication between groups, and having more people outside the testing community address the social and technical values of testing.

Tests can be used effectively by integrating them with learning and by having testers examine the use of tests in the real world, by placing the responsibility for guarding the moral aspects of tests and by involving non-testing groups in the discourse of testing. There is a need for greater awareness of and increased responsibility for uses and misuses of the devices which testers so naively develop. A move towards a meaningful dialogue (as opposed to inspection) may be one step in that direction.

REFERENCES

Foucault, M. 1979. *Discipline and Punish.* Vintage Books.

Fredriksen, J. and A. Collins. 1989. "A System Approach to Educational Testing." *Educational Researcher* 18, 27–32.

Freire, P. 1985. *The Politics of Education.* Bergin and Gravey Publishers, Inc. Massachusetts.

Madaus, G. 1991. *Current Trends in Testing in the USA.* Paper presented at the conference Testing and Evaluation: Feedback Strategies for Improvement of Foreign Language Learning, February 4-5. Washington D.C.: The National Foreign Language Center.

MacIntyre, A. 1984. *After Virtue* (2nd ed.). Notre Dame, IN: University of Notre Dame Press.

McLuhan, M. 1983. *Understanding Media: The Extensions of Man.* London: Routledge.

Nitko, A. 1989. Designing Tests That Are Integrated with Instruction. In R. Linn, Ed. *Educational Measurement.* Macmillan Publishing Company: New York.

Schwandt, A. 1989. Recapturing Moral Discourse in Evaluation. *Educational Researcher* 18, 8, 11–16.

Shepard, L. 1991. Psychometricians' Beliefs About Learning. *Educational Researcher* 20, 7, 2–9.

Shohamy, E. 1992. *The Power of Tests: A Study of the Impact of Three Language Tests on Teaching and Learning.* Paper presented at the Language Testing Research Colloquium, Vancouver, Canada, March 1992.

Stake, R. In press. *Effects on Changes in Assessment Policy.* Greenwich, CT: JAI Press.

INSTITUTIONAL AND POLITICAL ASPECTS OF LANGUAGE TESTING

Bernard Spolsky
Bar-Ilan University

The paper illustrates the non-theoretical influences on language testing that are so often ignored in histories written about the field. The writer argues that taking into account the institutional, social and economic situation can shed light on how new ideas on testing spread and how and why certain tests have become what they are now. The argument is illustrated by referring to two well-known test systems, the American TOEFL and the British Cambridge EFL examinations.

Most histories of language teaching and language testing have been written as though the driving force in the development of the fields has been advances in methodology and theory. We regularly talk and write (I know because I have done it) in terms of progress and periods. With this view of history, we have difficulty in understanding why a theoretical breakthrough (especially if it is one that we have just proposed) does not immediately win absolute acceptance and total implementation. Only recently (e.g. Richards 1984, Pennycook 1989, Phillipson 1992) has our attention started to be drawn to external, non-theoretical, institutional, social forces that on deeper analysis often turn out to be much more powerful explanations of actual practice. We forget for instance the enormous power of institutional inertia; as any one who has ever sat on a university committee will agree, it is much easier to think up reasons against a change than to provide arguments in its support. There are of course important and valuable reasons why this is so. Institutions are explainable as much through their history as through the logic of their present operation; they function because of their constancy, their imperviousness to other than minor changes. It is much easier to come up with a new theory than to find a way of fitting its implementation into an existing institution.

A clearer view of the history of the field will emerge once we are willing to look carefully not just at the ideas that underlie it, but also the institutional, social and economic situation in which they are realized. For instance, the development of language laboratories appears to have been an offshoot of developments of technology rather than an answer to theoretical needs; the enormous growth of the demand for English language teaching throughout the world is a reason of the huge language teaching industry and related publishing (and testing) businesses, where new theories are of interest as sales pitches; the move towards a European economic community clarified, as no theoretical

approach would have, the requirement to define language teaching goals as precisely as did the notional-functional syllabus.

The development of English language tests seemed to me an obvious field to start off this wider study. Language testing is of central interest because of the various factions involved. One of the reasons for my continuing fascination with the field is the fact that it constantly sets practical and theoretical issues into fruitful tension: the needs of the tester regularly challenge the theorist just as the findings of the theorist repeatedly tempt the tester. While it is fairly easy to come up with new assessment procedures, it remains difficult to explain exactly what is being measured, a situation that guarantees a continuing productive stress. As if this first cause of tension were not enough, there is a second one provided by the fact that at least two disciplines have proprietary interest in the theory behind language testing: both language learning theorists and measurement experts have their own independent (and perhaps unresolvable) notions of what is involved.

In my first excursion into language testing history (Spolsky 1977), I proposed a distinction that is now widely accepted between three periods or approaches to language tested, which I now label *traditional*, *modern*, and *post-modern*. The traditional approach assumed the full authority of an "examiner" to judge intuitively and accurately the spoken or written performance of a candidate. The modern approach questioned the reliability of such judgments, and called for psychometrically reliable and valid tests, with explicit criteria. Post-modern testing aims to find a blend: to test humanly (as in the traditional approach) the authentic language use of the candidate (the communicative competence rather than knowledge of just some limited discrete items) but to do it fairly (reliably) and validly. The wide usefulness of the distinction suggests there is much truth in it.

My reentry into the study of language testing history was in part an attempt to see the accuracy of the claims I had made a decade early, but more precisely it had two distinct but related motivations. It arose in part out of an interest in understanding why two highly successful batteries of tests of English as a second or foreign language seem to be so different in surface appearance and underlying ideology. The second and more general question that the study promised to help me start to understand was the course and nature of innovation in applied linguistics: the way in which developments in theory are blended with the requirements and possibilities of real-life implementation.

The first interest was related to the question of why the Test of English as a Foreign Language administered world-wide by the Princeton-based Educational Testing Service is so distinct from the examinations for the Certificate of Proficiency in English and the First Certificate in English offered by the University of Cambridge Local Examinations Syndicate. This question came up when I was invited to serve as a member of the advisory committee of a study of the comparability of these two tests conducted for UCLES under the direction

of Lyle Bachman. Bachman and his team had developed interesting and valuable approaches to describe the differences between the two sets of tests, and examined the possibility of establishing formal comparability. Their work showed the extent of the existing differences, and the interesting ways that each testing agency is now modifying its own approach, with the possibility of an ultimate ideological rapprochement.

In the course of meetings of the advisory committee which took place in 1988 and 1989, I was struck by the differences in language testing ideology expressed by British and American experts. To oversimplify, the British approach might be characterized as a continued belief in the traditional language examination and in the ability of the examiner, under the control of an elaborate but not fully articulated or explicit system of moderation, to give accurate impressionistic marks for the observed performance of candidates in writing or speaking, a continued suspicion of the objective test and a very cautious use of accompanying psychometric theory. In contrast, to oversimplify again, the American approach is based on fairly complete submission to the requirements of psychometric theory and objective testing, so that in much of the world objective tests are referred to as American tests, although American language tests have started to show a slowly shrinking reluctance to venture into the psychometrically dangerous waters of assessment of direct language performance.

The differences are striking; but what is equally striking are the similarities. First, one might mention that both the Educational Testing Service and the University of Cambridge Local Examinations Syndicate have developed very efficient and extremely profitable test batteries that clearly provide an excellent and satisfying service to all that use them. In 1986-7, TOEFL passed the 500,000 mark in the number of candidates taking the test annually, and numbers continue to grow. Similarly, over 200,000 candidates take UCLES English as a Foreign Language examinations each year. In each institution, one finds the same intriguing and fruitful tensions between institutional demands for continuity, and research interests in change; in each, one finds a strong professional commitment to excellence allied with a strong concern for continued profitability.

The second basic similarity is something else that blurs the differences: experts on both sides of the Atlantic show equal sophistication in their published work in various aspects of language assessment and testing in general; there are British experts and British tests, including for instance the IELTS test developed jointly by the Syndicate and the British Council, which demonstrates full mastery of psychometric principles, and indeed, much of the work in developing objective tests owes its impetus to research published by British scholars criticizing the traditional examination.

The similarities make the differences more interesting and perhaps harder to explain. Put more precisely, one question I want to try to answer is why did

TOEFL, as it was first developed, take the form of a purely objective, multiple-choice test, ready to fit easily (as it did after it had been developed) into the Educational Testing Service operation, while at the very same time, the University of Cambridge Local Examinations Syndicate English language examinations managed to resist the pressures, internal as well as external, calling for objective tests and do no more than add a couple of semi-objective sections.

A second question concerns the development of TOEFL itself, both before and after the move to Educational Testing Service, and that is how it managed to stay more or less independent of ongoing developments in the foreign language testing field. How is it, for instance, that the notions of oral testing that were by then so well established in the Cambridge tests, and closer at hand, in the Oral Interview of the Foreign Service Institute, had no influence on TOEFL until more than twenty years later? And how did TOEFL remain independent of the parallel foreign language tests, a situation that seems to have survived locating the two testing programs in neighboring offices?

These specific questions form part of a wider one concerning the nature of applied linguistics. Most histories of such aspects of applied linguistics as language teaching have been written as though the only influence in the development of the field has been advances in methodology and theory. We regularly talk and write (I know because I have done it) in terms of theory-dominated periods: we analyze the Audio-Lingual Method as the result of the application of structural linguistics and Skinnerian learning theory; we look at the cognitive approaches as following the theoretical revolutions of transformational- generative grammar; we consider the notional functional syllabus as related to theories of pragmatics and communicative competence. In our analysis, we have difficulty in understanding why a theoretical breakthrough (especially if it is one that we have just proposed) does not immediately gain complete acceptance and full implementation. Only occasionally is attention drawn to external, non-theoretical, institutional, social forces that on deeper analysis often turn out to be much more powerful explanations of actual practice. We forget for instance the enormous power of institutional inertia; as any one who has ever sat on a university committee will agree, it is much easier to think up reasons against a change than provide arguments in its support. There are of course important and valuable reasons why this is so. Institutions are explainable as much through their history as through the logic of their present operation; they function because of their constancy, their relative imperviousness to other than minor changes. It is much easier to come up with a new theory than to find a way of fitting its implementation into an existing institution.

A public language tests needs to satisfy several different sets of criteria, criteria drawn from different areas and representing the interests of several different groups. The language tester is responsible for and responsive to theory derived from two unrelated fields, linguistics and psychometrics, and at the same is

directed and constrained by practical institutional, economic, social and even political demands. Any large-scale test or examination battery will be the result of appropriate compromises between these theoretical and practical demands. By understanding how this has happened in the past, we will be in a sounder position to argue how we should proceed in the future.

I have so far published two studies which report on my current research into this question. The first to appear (Spolsky 1990) is a detailed recounting and analysis of the 1961 meeting that serves as the first steps in the development of TOEFL. At that meeting, there was a clear discussion of the state of the art in language testing, and a classic paper by John Carroll on the need for integrative tests. Two aspects of the meeting are important to our main points here today. The first concerns the motivations of the meeting. The other papers at the meeting make clear that the concerns of the convenors were practical: the papers by Alatis (1961) and Slocum (1961) present in detail the institutional and political pressures and the practical problems that were to be solved by a new test; the people invited were testers and test users as well as theorists. The second concerns the decisions made by the conference (and their later fate). While the two major batteries described at the meeting both included writing samples and oral interviews, and while the meeting is mainly remembered in the field for the paper in which Carroll (1961) rehearsed the arguments for integrative tests, the test battery that was developed in the next couple of years was as close as may be to a discrete point test, entirely multiple choice, with no free writing or speaking. At the meeting, there are signs, even clearer in the story of the next four years (Spolsky, in preparation), of how institutional concerns were to swamp theoretical judgments.

In the second of the studies I have so far published (Spolsky 1990), I describe a paper by John Roach (1945) that I found in a box in the archives of the University of Cambridge Local Examinations Syndicate. The paper dealt with research issues concerning the reliability of judgment in oral language examinations, issues that were not to appear in the testing literature for thirty years (Wilds 1975), although they were discussed internally at the Foreign Service Institute as early as the 1950s. Developed originally between 1952 and 1956, the first published account (Rice, 1959) gave only a brief description of the FSI test: fuller discussion of the nature and principles underlying the interview had to wait until the 1970s when papers by Wilds (1975) and Jones (1979) were commissioned, and a two-day conference was devoted to the topic (Clark 1978).

These various papers reveal how the FSI testers understood the special problems of oral language testing - to elicit a useful sample of speech, and then to arrive at a reliable and valid judgment of it - and the solutions that they proposed. The developers of the FSI considered themselves pioneers working in virgin territory: as Lowe (1988:16) put it, "Linguists at FSI would have gladly adopted or modified a system for ascertaining oral proficiency if one existed; the fact was that none did."

Had they seen Roach's report, they would have been much farther ahead. Reading the report with the benefit of forty-five years of further experience and fifteen years of discussion of tests of oral proficiency, one can see how Roach had managed to identify some of the major problems. First, it must be stressed that he was working within constraints that he understood both of technology (recording was possible but difficult) and institutional requirements: the conservatism of the examination syllabus, the expense of providing for joint examining. Second, he recognized the fundamental psychometric goals of reliability and validity. For the former, he proposed examiner training, joint examining, and guidance in what to judge. Validity for him was concerned mainly with the need to find a way of applying in examinations with multiple examiners the common standards set by the Syndicate. Here again, joint examining, specific efforts by examiners to agree on standards, the use of recorded samples, and the correlation of the more subjective oral marks with more objective parts of the examination (including dictation), were the system he proposed.

Roach (1945) then showed a good understanding of some of the basic questions in oral language testing, and offered a number of valuable solutions. The paper's influence however appears to have been quite limited, even within the University of Cambridge Local Examinations Syndicate, for Roach himself left its employ shortly after the report was completed. Thus, in 1963, when the Executive Committee in charge of the Syndicate's examinations in English for foreign students was considering a revision of the Proficiency syllabus, the issue of oral testing came up, but there was no reference to the Roach report; rather, a suggestion was made to ask John Trim to report on research in oral testing. The lack of external impact can also be understood. First, the report was published internally and distributed only among the Syndicate and its examiners. While a second version was in fact offered for sale, and while there is some indication of wider circulation, the lack of citation in language testing literature is evidence of its limited impact. None of the standard books on language testing, whether published in the US or Great Britain, cite Roach (1945) or appear to be aware of his report.

These two examples already suggest some of the benefits that accrue from a wider view of our field and its history. Language testing takes place in a political and an institutional context; the political and institutional forces will necessarily have a great influence on its development. Valuable as a history of ideas may be, we need more solidly based histories to understand how our field works.

REFERENCES

Alatis, J. E. 1961. Testing the English language proficiency of foreign students. In *Testing the English proficiency of foreign students* (pp. 11-29). Washington, DC: Center for Applied Linguistics.

Carroll, J. B. 1961. Fundamental considerations in testing for English language proficiency of foreign students. In *Testing the English proficiency of foreign students* (pp. 30-40). Washington, DC: Center for Applied Linguistics.

Clark, J. L. D. (Ed.). 1978. *Direct testing of speaking proficiency: Theory and application*. Princeton, NJ: Educational Testing Service.

Jones, R. L. 1979. The oral interview of the Foreign Service Institute. In B. Spolsky (Ed.), *Some major tests* (pp. 104-5). Washington, DC: Center for Applied Linguistics.

Lowe, P., Jr. 1988. The unassimilated history. In P. Lowe, Jr. & C. W. Stansfield (Eds.), *Second language proficiency assessment: Current issues*. Englewood Cliffs, NJ: Prentice Hall Regents.

Pennycook, A. 1989. The concept of method, interested knowledge, and the politics of language teaching. *TESOL Quarterly*, 23(4), 589-619.

Phillipson, R. 1992. *Linguistic imperialism* Oxford: Oxford University Press..

Richards, J. C. 1984. The secret life of methods. *TESOL Quarterly*, 18(1), 7-23.

Roach, J. O. 1945. *Some problems of oral examinations in modern languages: An experimental approach based on the Cambridge Examinations in English for foreign students, being a report circulated to oral examiners and local examiners for those examinations*. Cambridge: Local Examinations Syndicate.

Slocum, J. B. 1961. Assessing the English proficiency of foreign students: Point of view of US universities. In *Testing the English proficiency of foreign students* (pp. 99-102). Washington, DC: Center for Applied Linguistics.

Spolsky, B. 1977. Language testing: Art or science. In G. Nickel (Ed.), *Proceedings of the Fourth International Congress f Applied Linguistics* (Vol. 3, pp. 7-28). Stuttgart: Hochschulverlag.

Spolsky, B. 1990a. Oral examinations: An historical note. *Language Testing*, 7(2), 158-173.

Spolsky, B. 1990b. The prehistory of TOEFL. *Language Testing*, 7(1), 98-118.

Spolsky, B. (in preparation). *Measured words: a history of modern language testing up to 1965*.

Wilds, C. 1975. The oral interview test. In B. Spolsky & R. L. Jones (Eds.), *Testing language proficiency* (pp. 29-37). Washington, DC: Center for Applied Linguistics.

LANGUAGE PLANNING POLICY AND DEVELOPMENT OF FL PROFICIENCY IN FINLAND

Sauli Takala
University of Jyväskylä
Institute for Educational Research

The article gives first a brief outline of the official language policy of Finland and gives an account of the recent extensive work on the definition of the national language teaching policy. The plans are characterized by considerable safeguards for linguistic minorities and systematic attempts at diversifying both language teaching and learning. It is shown that extension of language teaching has brought about positive developments in language learning. It is also demonstrated that L1 can have a strong impact on the level of L2 learning.

1. SOME HISTORY

Finland became part of the Swedish kingdom some 900 years ago and remained its eastern province until 1809, when Russia became the new master due to the Napoleonic wars. Swedish was the official language of government. The school ordinance of 1649 made Swedish the language of instruction (in addition to Latin). In the 1700's, Swedish became the sole language of instruction in secondary education. Finnish obtained a recognized position only slowly. Below are some milestones in the development of Finnish language and culture.

* The first books in Finnish were published in the 1540's.
* Finnish became an elective school subject in secondary schools in 1841.
* The first Finnish-medium secondary school founded in Jyväskylä in 1858.
* The first Finnish-medium teacher training seminar founded in Jyväskylä in 1863.
* First professorship of Finnish was established at the University of Helsinki in 1851.
* The University of Helsinki was made bilingual in 1923, Finnish the main language of instruction in 1937.

The Swedish-speaking population has always been a minority in Finland. The proportion of Swedish-speaking persons was about 17% in 1600 (70,000 in all), reaching highest absolute numbers in 1940 (355,000, 9,5%) and decreasing to

300,000 in 1980 (about 6,3%). Statistics about the number of Lapps in Finland vary according to what statistics are used. They vary from a few thousand to a few dozen thousand.

2. OFFICIAL LANGUAGE POLICY

The Constitution of the new republic (1919) made Finland officially bilingual, guarantees basic language rights and ensures that equal consideration shall be given to satisfying the material and cultural needs of both language groups (i.e., Finnish speakers and Swedish speakers). The linguistic status of each local unit of self-government (municipality) is determined by the Language Act of 1922, amended in 1975 in the interest of the Swedish-speaking minority. The municipality is bilingual if the minority amounts to 8% of the total population in the municipality or is at least 3,000 people. The status of each municipality is assessed every ten years, after the national census. This was done last in 1982 and covers the period 1983-1992. A municipality cannot be declared monolingual unless the share of the minority has dropped below 6% and the Government can grant special dispensation for ten years even if that figure has not been reached. The Swedish-speaking population on the Åland Islands has been granted strong unilingual guarantees.

There are 396 unilingually Finnish municipalities, 21 bilingual municipalities with Finnish majority, 20 bilingual municipalities with Swedish as the majority language, and 24 unilingually Swedish municipalities, 16 of which are on the Åland Islands. It is of extraordinary significance to the Swedish- speaking minority in Finland that most of them now live in bilingual municipalities.

The language rights of the Lapp minority have been officially ratified only this year (Act 516/91). From the beginning of 1992, the Lapps have full right to use the Same language in conducting official business. Same can, however, be used as a 'legitimated' language mainly only in the Same-speaking region in the north. Government and municipal officials are not required to be able to speak Same, since interpretation and translation can be used. A language bureau will be set up to look after such services.

Pupils living in the Same-speaking region have the right to get instruction in Same both as a first and second language. Same can also be used as a medium of instruction. From 1980, it has been possible to take an examination in Same in the matriculation examination as an optional language test. A new statute (14.6.1991/934) has made it possible to have Same as the first language in the senior secondary school, which means that Same can soon be taken as a mother tongue test in the matriculation examination.

3. LANGUAGE TEACHING POLICY

Until the end of the 1960's second and foreign language teaching was limited to the junior and senior secondary schools. The introduction of the 9-year comprehensive school at the beginning of the 1970's brought about a drastic change: all pupils study the second official language of the country and one foreign language as compulsory subjects (A-language starts in grade 3 at age 9, B-language in grade 7 at age 13). Pupils can also read a second foreign language as an optional subject in grades 8-9.

English and Swedish were prioritized as the A-language, since it needed a Cabinet permission to teach also French, German or Russian. The local community could decide which of the two languages, English or Swedish, was to be the main or only language taught. If there was a choice, pupils and their parents could freely decide on which language to choose as the A- language. English has long dominated as the A-language among Finnish-speaking pupils, which means that Swedish has been the main B-language. German has always dominated as the C-language, and girls have constituted about two-thirds of C-language readers.

The dominance of English has worried the national authorities, and several commissions and working groups have tried to find ways of diversifying the study of the A-language. A national commission that reported in 1979 made several recommendations to that effect.

The commission proposed that communities should be obliged to offer a range of choice which depended on the size of the community:

- 100,000 or more inhabitants: a choice of five languages must be offered (the other official language of the country, English, French, German and Russian)
- at least 50,000 inhabitants: a choice of four languages
- at least 20,000 inhabitants: a choice of three languages.

This was approved, and the provincial governments were required to prepare a regional plan and the municipalities to prepare a local language teaching plan to implement the requirements. Thus the planning of language teaching policy was extended to cover all three levels: national, regional and local (see e.g. Nikki 1992).

The commission also proposed some quantitative targets concerning the proportion of people who should know various languages in the future. It also suggested target figures for the study of various languages as the A-language. The Cabinet did set such targets for 1985-1991.

As it was evident that the targets were not met, another working group recommended that it should be possible for pupils to begin studying English already in grade 5 (instead of grade 7), if they chose French, German or Russian as their A-language. Using English as a 'carrot' was believed to make the other languages more attractive: pupils could learn two world languages well. Pilot studies showed that choices became more balanced. Recently, the so-called E-English has become a regular part of the national language teaching policy.

General plan for language diversification by the Cabinet for 1985-1991 vs. its implementation, A-language in the comprehensive school

A-language	Plan	Actual/1987	Helsinki/1990 after 'E-English'
English	81.5%	86.8%	59.2%
Swedish	9.0%	6.0%	21.6%
Finnish	4.5%	4.8%	
French	1.5%	0.7%	6.2%
German	2.0%	1.0%	11.1%
Russian	1.5%	0.6%	1.9%

General plan for language diversification by the Cabinet for 1985-1991 vs. its implementation, C-language in the comprehensive school

C-language	Plan	Actual/1987
English	3.0%	1.1%
French	7.0%	8.7%
German	21.0%	25.9%
Russian	4.0%	2.3%
Other	0.3%	0.5%
% of age group	35.3%	38.5%

General plan for language diversification by the Cabinet for 1985-1991 vs. its implementation, C- and D- languages in the senior secondary school

C/D-language	Plan	Actual/1987
English	3.0%	1.7%
French	20.0%	19.4%
German	62.0%	53.2%
Russian	10.0%	5.9%
Spanish	1.0%	0.6%
Other	1.5%	0.8%
% of all	97.5%	81.6%
Late 70's	125.0%	

A new commission quite recently proposed that it should also be possible to study Spanish as an A-language, and Italian as a C-language. This would be a major change since, at present, those languages can be studied first in the senior

secondary school as a D-language. Quite recently, all restrictions concerning the C and D languages have been removed.

As proposed by another recent working party, it has now become possible for schools to teach content-area subjects using a second or foreign language. Pilot schemes are underway using both total and partial, immersion approaches. The University of Jyväskylä plays a key role in providing PD training for teachers.

In the senior secondary school, it was customary to study the other official language of the country and two foreign languages. If you had chosen a language line, you would study one further language. A few years ago, the study of the second foreign language was made optional for those students who take an advanced course in mathematics. This has brought about a distinct drop in the study of foreign languages in the academic secondary school. This is shown in the matriculation examination in C/D-languages: a drop from 76% to 38% in ten years.

By contrast, in the 1970's the study of the other official language of the country and one foreign language were introduced as *compulsory subjects* to all students who transferred to the vocational schools or institutes after the comprehensive school.

When the university degrees were reformed in the 1970's, the study of foreign languages and of the other official language of the country was made *obligatory for all college and university students*. Language centers were set up to handle the teaching and a coordinating centre was set up at the University of Jyväskylä.

The latest development in language planning concerns timetables: the government has established minimum thresholds for lessons in school subjects, including languages. Swedish was retained as an obligatory subject for all Finnish-speaking pupils but its lesson quota was reduced.

The A-language will have at least 8 lessons both during the lower stage and the upper stage of the comprehensive school (earlier 15-19 lessons), the B-language at least 6 lessons (earlier 9 lessons), the optional E-language 4 lessons in the lower stage (earlier 4 lessons), the C-language not specified (earlier 5 lessons). In the upper secondary school the A-language will get at least 6 courses (earlier 8), the B-language at least 5 courses (earlier 7) and the other optional languages may have a total of 16 courses combined.

4. SOME EMPIRICAL RESULTS ON LANGUAGE PROFICIENCY

Language proficiency has been monitored in Finland since the comprehensive school was introduced in the early 1970's. The more than one hundred years old matriculation examination, which is currently taken annually by one half of the

whole age group, makes it also possible to carry out some interesting comparisons. I will first briefly summarize the main findings about FL achievement in the comprehensive school.

4.1. Foreign language achievement in the comprehensive school

It can be said without much exaggeration that language testing started to become professional in Finland when the experimentation with the comprehensive school started in the late 1960's. Since the knowledge of foreign languages has always been very important in the secondary school and since foreign languages make up an important part of the matriculation examination, it was important to assess the level of achievement in the experimental comprehensive schools and compare it with the level of achievement in the well-established secondary school. Table 1 summarizes some of the comparison data.

TABLE 1. Mean percent correct figures based on identical items in the 1971 and 1979 assessments; whole age groups and broken down by strata, pupils aged 14-15.

	IEA 1971	Nat. Ass. 1979	IEA Sec. Sch.	IEA Elem. Sch.	National Assessment Set A	Set B	Set C
List. compr.	40.3%	59.4%	65.7%	15.0%	80.1%	55.2%	33.3%
Read. compr.	40.4%	62.7%	63.0%	18.0%	81.7%	56.7%	42.1%

The results indicate that when we look at the "national yield" in English proficiency (the first two columns) the introduction of the comprehensive school has meant a 20 percentage point increase in listening and reading comprehension attainment from 1971 to 1979. We also note that the pupils in Set A (the most able pupils) had achieved the best result and that Set B pupils did also quite well in comparison to the secondary school pupils. The level of attainment in the higher grades of the elementary school was very modest in the early 1970's. This is understandable because foreign languages were just beginning to make their way to the former elementary school. The level of performance of the weakest comprehensive school pupils (Set C) was quite low as well.

The second national assessment in 1991, which in English focussed on listening comprehension, indicated (Havola & Saari 1993) that ninth graders in the comprehensive school (representing the whole age group) did about equally well in the identical tests as the pre-university students (21% of the age group) twenty years earlier. Obviously, there are several reasons that account for this outcome,

some of which have nothing to do with formal teaching in schools, but it does represent a genuine improvement of understanding spoken English in Finland.

The author's Ph.D. thesis (Takala 1984) estimated the comprehensive school pupils' vocabulary size in English in 1979 after nine years of English (with typically 2 lessons a week). The results indicated that Set A pupils had learned about 1500 words, Set B pupils about 1000 words and Set C pupils about 450 words. Related to the total number of lessons, this means that Set A pupils had learned 3-5 words per lesson, Set B pupils about 2 words and Set C pupils about one word per lesson. The results also showed that the pupils' word formation skills were poorly developed. The overall outcome cannot be considered very satisfactory, especially since similar results were also obtained in Swedish, which is an obligatory language for all Finnish-speaking pupils.

Since the results from the 1979 assessment were published, development work on vocabulary teaching and learning was initiated, which led to a report edited by the author (Takala 1989). A second national assessment carried out in 1991 showed, in fact, that whereas only about 60% of core vocabulary in Swedish was known in 1979 the corresponding figure for 1991 was about 80% (Kärkkäinen 1993).

The national assessments in Finland always seek to measure "affective outcomes" as well. The results (Kärkkäinen 1993) show that the rank of English in terms of importance rose from three to top the list. In terms of pleasantness, English rose from rank 10 to rank 5. Swedish displays an interesting pattern: it has slipped from rank 4 to 7 in importance whereas its rank in terms of pleasantness has been much lower and has shown a declining trend: rank 12 in 1979, rank 16 in 1991.

4.2. Foreign language achievement in the upper secondary school

The matriculation examination is taken by about half of the age group each spring. It includes a compulsory test in one foreign language (mostly English) and in the second official language of the country (Swedish for the majority of the students). The foreign language tests are identical for Finnish-speaking and Swedish-speaking students, which makes it possible to compare their performance level (Table 2).

The results indicate that Swedish-speaking students do better in English tests than Finnish-speaking students. In the MC listening comprehension test the percentage point difference varies between 7.2 and 13.9, in the MC reading comprehension test between 5.5 and 9.9, and in the MC cloze integrated test of grammar and vocabulary test between 0.2 and 14.0. The cloze test was introduced in 1990 and during the first two years the difference between the two groups was negligible but has since grown and is similar to the difference in the LC and RC tests. It should be noted that the Swedish-speaking students

have studied English two years less and spent less time on studying English (see Ringbom, 1987, p.81, for earlier statistics with the same trend).

TABLE 2. Average per cent correct values in multiple choice English tests, Finnish- and Swedish-speaking students.

Year	Finnish-speaking			Swedish-speaking		
	Listening	Reading	Cloze	Listening	Reading	Cloze
Spring 1988	70.2%	78.9%	-	81.4%	86.4%	-
Spring 1989	74.9%	72.4%	-	88.8%	81.8%	-
Spring 1990	76.5%	86.3%	71.8%	86.5%	91.8%	72.0%
Spring 1991	76.8%	75.9%	69.8%	87.3%	81.8%	72.0%
Spring 1992	78.2%	79.3%	72.8%	89.3%	86.8%	86.8%
Spring 1993	80.0%	71.2%	77.2%	87.2%	79.2%	87.6%

Swedish and English are both Germanic languages, which show a fair amount of similarity in terms of underlying grammatical structure and cognate words. Finnish, by contrast, belongs to an entirely different language family. Since there are no major cultural or other contextual differences between the Finnish-speaking and Swedish-speaking populations and schools, the most probable reason for the difference in performance can be ascribed to the influence of the first language in foreign language learning. The first language can powerfully influence the rate and level of foreign language learning. The same trend has been found in the C/D language (shorter language courses) tests in German and French.

5. OUTLOOK

There is a strong trend in Finland at all levels of the educational system to remove centrally prescribed regulations and to delegate decision making to schools and allow greater freedom of choice for pupils. This means that there will be a need to monitor what changes will take place in foreign language study in Finland. How many pupils/students will study what languages for how long and what have they learned?

Like all universities, the University of Jyväskylä has been required to indicate its areas of current and emerging strength. These plans are expected to help universities to plan their resource allocation better. One of the *university-level* areas of strength in Jyväskylä is "research on language learning, teaching and communication". One of the priority areas in the *educational field* in Jyväskylä is "evaluation of educational productivity". These two combined will mean that the University of Jyväskylä will continue to play a central role in language testing and evaluation in Finland. One of the future activities will be

participation in the IEA comparative study of foreign language learning, which is expected to be implemented around the mid-1990's.

REFERENCES

Havola, L. & Saari, H. 1993. Kaksikymmentä vuotta englannin kuullun ymmärtämistä - onko taidoissa edistytty [Twenty years of English listening comprehension in Finland - Has there been any improvement?]. In P. Linnakylä & H. Saari (Eds.) *Oppiiko oppilas peruskoulussa?* University of Jyväskylä. Institute for Educational Research, 157-173.

Kärkkäinen, K. 1993. Ruotsiko ongelma? [Swedish a problem?] In P. Linnakylä & H. Saari (Eds.) *Oppiiko oppilas peruskoulussa?* University of Jyväskylä. Institute for Educational Research, 121-139.

Nikki, M.-L. 1992. *The implementation of the Finnish national plan for foreign language teaching.* University of Jyväskylä. Jyväskylä Studies in Education, Psychology and Social Research Nr 87.

Ringbom, H. 1987. *The Role of the First Language in Foreign Language Learning.* Clevedon: Multilingual Matters.

Takala, S. 1984. *Evaluation of Students' Knowledge of English Vocabulary in the Finnish Comprehensive School.* University of Jyväskylä. Institute for Educational Research. Research Report Nr 350.

Takala, S. 1989. *Sanaston opettaminen ja oppiminen [Vocabulary instruction and learning].* University of Jyväskylä. Institute for Educational Research. Theory into Practice Nr 44.

LANGUAGE POLICY AND LANGUAGE TESTING

Ingrid Wijgh
Cito
Arnhem, Holland

The article deals with the introduction of the first common curriculum for basic secondary education in the Netherlands. The learning objectives for foreign languages are described, as well as the plans to test their achievement. The writer also discusses the experimental tests aimed at measuring the objectives which were grouped into larger categories such as communicative ability (reading, listening, speaking, writing), socio-cultural competence and compensatory strategies.

1. EDUCATIONAL POLICY

The first stage of secondary education has been the subject of continuous concern for over fifteen years in Dutch educational policy. The main concern has been and still is, to postpone the definite choice pupils have to make and, in doing so, to avoid premature selection. All this in view of the slogan of the late sixties: equal opportunities for everyone. After finishing primary education at the age of 12, Dutch children have to make their choice between four types of secondary education to continue their studies: Lbo, Mavo, Havo or Vwo.

Lbo stands for Junior Secondary Vocational Education. This is a four year vocational training with a great deal of general subject matters among which one foreign language and sometimes a second one. Mavo, Junior General Secondary Education, offers courses in three foreign languages. Pupils are entitled to drop one or two languages, but one foreign language is compulsory until the end of the four year school period. Havo, Senior General Secondary Education, takes five years, pupils follow courses in three foreign languages. Vwo, Pre University Education, also offers three foreign languages. This type of education takes six years.

This early choice and pre-selection at the age of 12 has for a long time been resented as undesirable by policy makers. Ever since 1975 every self-respecting government has proposed a new plan to reform and harmonize the first stage. That makes four plans, the last of which appeared in 1987. The times had been changing and so had the plans: the first one, which had the status of a working paper, stressed the general educational goals and proposed a comprehensive

school in which all pupils should stay together in heterogeneous groups for four years.

The last plan is the least ideologically inspired and the most practical and concrete of them all. However, the most recent proposals are based on a thorough analysis of Dutch society, its demographic and economic characteristics and its present and future needs. In the previous plans attempts were made to reduce the number of hours spent on foreign languages; the last one pays more attention to foreign language learning. This phenomenon is undoubtedly caused by the prospect of Europe 1992, when formal frontiers will be abolished in the European community. That expectation stimulates the commercial greediness; the Dutch, as so many others, expect to conquer enormous new markets and mastery of foreign languages is believed to constitute a precious instrument in the realization of that intention. Everywhere in the common market one can see a growing interest in foreign language learning.

The last plan for educational reform, which proposes a 'basic education' has come further than its predecessors. The bill proposing this reform has been discussed in Parliament in 1991 and it has been accepted by a large majority. Not without problems, as a matter of fact. The old battle between supporters and opponents of the comprehensive school was revived for a while. Arguments pro and contra the heterogeneous class were heard again. Now that the bill has been passed, for some a dream will come true, even if in a form very different from the original plan which goes back to the mid seventies. For others educational reality will turn into a nightmare. They argue that a uniform program for everyone is a spilling of talents, because the program will be trivial for the bright ones.

2. BASIC EDUCATION

In 1993 all pupils leaving primary school (at the age of twelve) will receive 'basic education', a common core curriculum, during a period of approximately three years. In that period they will have to study fifteen subjects: mathematics, biology, physics, chemistry, economics, home economics-health care, technology, history, geography, informatics, arts, physical education, Dutch, English and French or German.

This basic education does not affect the structure of the existing educational system; it just means a change in the curriculum. In point of fact, basic education means the introduction of a common curriculum, because earlier the aims of the first stage of secondary education had never been clearly described. They could only be derived from the methods used during that period and from the final examinations at the end of the second stage.

The general aims of basic education are:

- postponement of definite choice;
- raising the general level of achievement;
- reinforcement of the common cultural basis;
- encouraging equal opportunities for girls and boys.

Furthermore, specific objectives have been formulated for each of the fifteen subjects. These objectives are described in terms of pupils' behavior: "pupils are able to" etc. The specific objectives serve a double purpose; on the one hand they provide a functional set of knowledge and skills, ready for use, and on the other hand they provide a basis for further study. This difference is shown in context-dependent and context-independent objectives. 'The learners can order something' is an example of the former. 'The learners can understand the main idea of written texts' is an example of the latter.

In the history of Dutch educational policy the formulation of national specific objectives constitutes a memorable event. The Netherlands have had a long tradition of freedom of education; schools have the right to devise their own curriculum. This freedom of education is even laid down in the Constitution (1848) and implies everyone's freedom to provide education. As soon as national governments attempted to formulate educational objectives in more than a very general way, this provoked a lot of protest and was always considered an attack on the freedom of education. The specific objectives for the different subjects of basic education have provoked similar reactions. Nevertheless, the educational field has embraced the specific objectives with more or less enthusiasm depending on the various subjects and it appears politicians have done the same, at last.

3. FOREIGN LANGUAGE LEARNING IN BASIC EDUCATION

The formulation of the objectives for the various subjects had been assigned to groups of experts on behalf of the Minister of Education. In the case of foreign languages there was only one group for the three languages that have to be studied during basic education. Its chairman was Jan van Ek, the author of the well-known Threshold Level, which was published by the Council of Europe in 1975 (A revised and extended version was published in 1991). Members of the group were recruited from the Institute for Curriculum Development, the Institute of Educational Measurement, the professional organization of foreign language teachers, teacher training centers and educational publishers.

In 1989 the group presented a set of educational objectives for English, German and French. No distinction had been made between first, second or third foreign language: the aims were identical for the three foreign languages concerned and constituted a minimum level of proficiency. Differentiations between those three

rites and conventions, as far as they are manifest in the communicative behavior of the members of the other society.

The orientation in foreign language learning is a domain of a slightly different order. Here language learning is considered in its broad educational context. Learners are expected to have insight into their own objectives, to know which procedures and aids will contribute to achieving those objectives and to assessing their progress in foreign language acquisition. In doing so, they are relating the learning of foreign languages to their own needs and experiences.

4. ASSESSMENT IN BASIC EDUCATION

The Basic Education Bill makes some general remarks about evaluation. The attainment of the objectives has to be assessed by means of national tests, provided by the Minister of Education. Schools will be relatively free to determine the specific times at which the tests are to be taken by their pupils. They will also be free to decide whether pupils will have succeeded in attaining the educational targets, but they will be obliged to administer the tests provided by public authority. Thus the government will perform its duty of guaranteeing the quality of education.

Since 1990, Cito, the National Institute of Educational Measurement, has been working on the testing component of basic education. Cito has put forward the outlines of a flexible and school-centered evaluation system. A system, which disturbs school routine as little as possible and fits into the daily school life. These characteristics impose certain constraints on test development. For instance, the taking of a test or a part of a test should not exceed the time limits of a regular lesson period. However, the ultimate decision on the assessment system has yet to be taken by the politicians. Though reforms in education almost always bring with it certain political problems, to professional test constructors like those working at Cito they signify a new challenge to try and find assessment formats appropriate to the objectives of basic education.

5. HOW TO TEST OBJECTIVES?

Hitherto experimental tests have been developed for all fifteen different subject matters of basic education, including French, German and English. Since objectives have been formulated under the guidance of Jan van Ek, it is not surprising that the specific objectives for foreign languages have been derived from the overall goal of communicative ability. For at least ten years now a lot of testing has been going on inspired by the concept of communicative competence. So the experimental language tests do not just arise out of

languages for Dutch learners should be made on the most concrete level: in the choice of learning materials like texts for reading and listening and in the actual test construction.

The general aim of language learning was formulated as follows:

> "In various communication situations, both in their own country and elsewhere, learners can use the foreign language in such a way as to satisfy those communication needs which, generally, are the most essential ones in these situations. The relevant situations have the following characteristics:

- they are of frequent occurrence and/or of (potential) importance to the learners;
- use of the foreign language is either required or advisable;
- they do not require more than a basic level of command of the foreign language, to be further characterized hereafter;
- they may involve the use of spoken language as well as written language, both receptively and productively."

Nineteen specific objectives (see Appendix) were derived from this overall goal and they were categorized in four domains and some sub-domains.

- communicative ability
  - . reading
  - . listening
  - . speaking
  - . writing
- compensatory strategies and techniques
- socio-cultural competence
- orientation in language learning

In the domain of the communicative ability all the objectives are specified according to the four skills. Reading and listening are considered the most important ones, just as speaking. Relatively little attention is paid to the writing skill.

Learners handle compensatory strategies and techniques in dealing with the demands of communication situations that they are not fully prepared for. This means for instance the proper use of a dictionary, the ability to deduce the meaning of unknown elements from a given context and also the ability to ask for explanation and repetition.

A learner is said to be socio-culturally competent when having some familiarity with everyday life, living conditions, interpersonal relations and major values and attitudes in the other country. Besides, he should have insight into social

nowhere. Test construction focusses on the domain of the communicative language ability and within that domain on the four skills.

These experimental tests have several functions;

- they are the operationalization of the specific objectives, formulated at a so-called intermediate level of abstraction. This facilitates communication about education between those involved in education, such as parents, teachers and curriculum developers.

- furthermore, they enable the legitimation of the test constructor's interpretation of those objectives by experts.
- they form one of the most concrete illustrations of the objectives and can thus play a role in the introduction of basic education. To teachers, they clarify the targets their pupils should have attained at the end of basic education.
- they can help to get information on the possible gap between the level aimed at by the test constructors and the actual level attained by the pupils at the end of basic education. Until now, the level is still an academic one: will pupils really be able to attain these targets after some 240 hours of foreign language courses? This has to be verified by empirical data. The experimental tests are the instruments to collect these data.

5.1. Reading skills

Tests of reading have been developed for French, German and English. They consist of approximately 25 open and closed questions on a number of authentic texts such as leaflets, news items, signs, announcements, advertisements, etc. Most of the texts are rather short, except one, which is a fairly long magazine text. One, sometimes two or three questions are asked on the short texts; between seven and ten questions on the long texts.

Those questions aim to represent objectives like 'identifying relevant information in functional texts', 'comparing information and drawing conclusions' and 'understanding the essence of texts'. The German and the French specimens have been pretested on a small scale. This pretest was meant to give answers on questions about the level of basic education and the degree of difficulty of the tests for the different groups of participants, i.e. pupils of lower vocational education, junior -, senior - and pre university education. Results are such that the idea of a one and only test for this very mixed ability group is not to be rejected off-hand. The differences in degree of difficulty are not dramatic: for pre university pupils the test is not trivial and for vocational pupils the test is not too difficult.

5.2. Listening skills

In the early nineties, the financial aspects of education and also of testing, have become more and more important in the Netherlands, as they have in so many other countries. Hence, test construction has to be more and more efficient. This means that more attention is being paid to the development of prototypes, to the recycling of existing tests and to a broader distribution of standardized tests.

Regular tests of listening comprehension, produced at Cito for the intermediate level of vocational education, seem to correspond with the objectives formulated for basic education such as 'identifying and understanding information on radio, television or by telephone' and 'understanding the substance of spoken texts'. So we are attempting to do some recycling. In the near future a research project will be set up to find out if this existing test of listening comprehension is suitable for the different groups of pupils receiving basic education. The results of this test administered to groups of pupils after either three years or two years of foreign language training will be examined. This is because, though originally meant to take three years, basic education may take between two and four years. Pre-university pupils are expected to finish their basic education, i.e. to attain the objectives, after two years of foreign language training. So there is some kind of tempo differentiation introduced into basic education.

5.3. Writing skills

The objectives in this field are rather minimal: emphasis is given to the receptive skills and the speaking proficiency. At the end of basic education pupils have to be able to fill in simple forms and to write short notes and letters. The assessment concentrates on aspects such as appropriateness and communicative value, rather than on grammatical correctness. This aspect becomes only important if it causes communication to break down.

5.4. Conversation skills

The objectives for speaking consist of 5 sub-domains: understanding, socializing, communication strategies (discourse- and compensatory strategies), functioning in daily-life situations and talking about certain topics. Test construction in this field means looking for the right format(s).

It might be efficient to test communication strategies and socializing in separate tests. Or it could be better to integrate those sub-domains within a structured interview and in role playing. The main problem is, besides reliable assessment, to develop elicitation techniques for 'functional language use', i.e. the right word at the right time in the right place. Most existing speaking tests, although claiming to test functions such as requesting, convincing and apologizing, can in fact be seen merely to test 'imparting and seeking factual information'. A

special research project has been set up to the development of speaking tests that will enable the assessment of functional language use.

5.5. Compensatory strategies, socio-cultural competence

The testing of these two domains has been integrated in the tests for communicative ability. The reading test contains questions such as "What would the word 'requin' mean?" This is asked after pupils have read a French article on a boy being attacked by a shark and saved by a dolphin (it is very improbable that they actually know the word 'requin'). In the speaking tests, pupils are supposed to ask for repetition whenever they do not understand their partner. One of the criteria when selecting appropriate texts for reading comprehension is their 'amount' of cultural information, i.e., do they describe aspects of daily life in another culture?

6. DISCUSSION

As mentioned above, test development for a brand-new curriculum is a challenging activity. Still there are three main problems to examine in the next years:

- Will it be possible to develop a one and only test for a very broad and heterogeneous public which is valid, reliable, not too difficult for the weaker pupils, nor too easy for the brighter ones?
- Will it turn out to be possible to test functional language use in a speaking test and will this test meet criteria of reliability and user's friendliness?
- What about basing decisions on test results? In the case of a pass/fail decision, how to determine the cutoff score? Or is it more appropriate in view of the aims of basic education to collect information by way of portfolio?

BIBLIOGRAPHY

Basisvorming in het onderwijs. 1986. Rapporten aan de regering, SDU, Den Haag.

Ek, J.A.van. 1986. *Objectives for foreign language learning. Volume I:Scope.* Council of Europe, Strasbourg.

Ek, J.A.van. 1975. *The Threshold Level in a European unit/credit system for modern language learning by adults.* Council of Europe, Strabourg.

Ek, J.A.van & J.L.M.Trim. 1991. *The Threshold Level 1990.* a revised and extended version of The threshold Level by J.A. van Ek, Council for cultural co-operation, Strasbourg.

Ministry of Education and Science. 1986. *Concept-Wetsvoorstel betreffende basisvorming.* MOW, Zoetermeer.

Ministry of Education and Science. 1989. *Advies over de voorlopige eindtermen basisvorming in het voortgezet onderwijs:moderne vreemde talen.* Den Haag.

Ministry of Education and Science. 1988. *The Dutch Education system.* Docinform no 332e, Zoetermeer.

Shohamy, E. A proposed framework for testing the oral language of second/foreign language learners. *SSLA* 10, pp.165-179.

The teaching of languages in the European community. 1988. Eurydice European Unit, Brussels.

The education structures in the member states of the European communities. 1986. Eurydice European Unit, Brussels.

Wijgh, I.F. 1990. *Afsluiting Basisvorming Moderne Vreemde Talen.* Cito, Arnhem.

Wijgh, I.F. 1990. *Over en sluiten.* de afsluiting van de basisvorming. Cito, Arnhem.

APPENDIX

Foreign language objectives

Basic Education (age 12-15)

General objective

In various communication situations, both in their own country and elsewhere, the learners can use the foreign language in such a way as to satisfy those communication needs which, generally, are the most essential ones in these situations. The relevant situations have the following characteristics:

- they are of frequent occurrence and/or of (potential) importance to the learners;
- use of the foreign language is either required or advisable;
- they do not require more than a basic level of command of the foreign language, to be further characterized hereafter;
- they may involve the use of spoken language as well as of written language, both receptively and productively.

A Communicative ability

Reading skills

1
The learners can understand the meaning of brief and simple instructions, notices, warnings and announcements, such as may be encountered:

- while travelling by car or public transport;
- on/in shops, department stores etc.;
- on/in public buildings, banks, post offices etc.;
- in hotels, youth hostels and camping-grounds.

2
The learners can identify and understand specific items of relevant information in functional texts, provided this information is formulated in simple terms. Examples of these texts are:
- leaflets and guides;
- 'what's on' -pages in newspapers, announcements of radio and TV programs;
- advertisements;

- weather forecasts;
- forms to be filled out (hotel registration etc.);
- timetables of public transport;
- confirmations of bookings;
- standard letters containing information about accommodation, fees and prices, admission times etc.;
- correspondence pages in magazines for young people.

3
The learners can compare various items of information from functional texts as meant under 2, and draw conclusions from the comparison.

4
The learners can understand the substance of written texts that are characterized by the following features:
- the structure of the text is simple and the vocabulary range is limited;
- the essential information is formulated explicitly.

In the selection of the texts the following four categories should be adequately represented:
- informative articles in newspapers and magazines about subjects within the experience of the learners;
- simple narrative, poetic or dramatic texts, provided they are close to the experience of the learners;
- personal correspondence on subjects within the following domains: private life, school, study, occupation, informal social contacts, leisure;
- texts that may provide insights into the daily life of those countries or areas where the foreign language is the native tongue.

5
The learners have gained experience in extensive reading, also of longer texts.

Listening skills

6
The learners can understand the meaning of simple and clearly spoken instructions, warnings and announcements, such as may be encountered while travelling by car or public transport. The spoken texts should, moreover, have the following features:
- a moderate, but not unnaturally low speed of delivery;
- an accent that is not strongly marked socially or regionally.

7
The learners can identify and understand specific items of information in functional texts that are clearly spoken on radio, television, or by telephone,

provided this information is formulated in simple terms. Examples of these texts are:

- weather forecasts;
- traffic information;
- information on medical services, opening hours, times of arrival and departure;
- brief newsflashes on subjects within the experience of the learners;
- programme announcements.

For restrictions on speed of delivery and accent: see under 6.

8

The learners can understand the substance of spoken texts that are characterized by the following features:

- the texts are fairly short;
- the structure of the text is simple and the vocabulary range is limited;
- the essential information is clearly and explicitly formulated.

In the selection of the texts the following three categories should be adequately represented:

- interviews and monologues concerning subjects within the domains indicated under 4 (personal correspondance);
- simple narrative, poetic or dramatic texts, provided they are close to the experience of the learners;
- texts that may provide insights into the daily life of those countries or areas where the foreign language is the native tongue.

For restrictions on speed of delivery and accent: see under 6.

Conversation skills

Understanding

9

The learners can understand their conversation partners in contexts as referred to under 'speaking', provided these partners are prepared to make allowances for the non-native speaker's limited command of the foreign language.

Speaking

10

Objective 10 is realized with regard to themes within the following domains:

- private life
- school, study, occupation
- informal social contacts
- leisure.

- The learners can give and obtain information.

- The learners can inquire about an opinion and express their own.
- The learners can describe somebody or something.
- The learners can talk about past events.
- The learners can talk about future events.
- The learners can express positive and negative feelings.
- The learners can make comparisons.
- The learners can express preferences.

Concerning objective 10:
The learners can express themselves in a manner appropriate to the communicative situations involved and in such a way that they can be understood without undue difficulty also by native speakers of the foreign language who have little or no experience with communication partners with a different language background. This implies a certain measure of correctness in the manner of expression, in the use of lexicon and in pronunciation, grammatical correctness playing a less prominent role except insofar as it is essential to effective communication.

The learners react to their conversation partners' questions and remarks with an understandable short sentence or single word. They can ask short understandable questions.

11
Objective 11 is realized in situational contexts such as:
- in shops
- at the post-office, at the bank
- in restaurants, cafés, snackbars
- at the police-station
- at the doctor's
- at the tourist-office
- on the road, travelling
- going out
- staying somewhere for the night

- The learners can ask/say what time it is.
- The learners can ask/show the way.
- The learners can buy something.
- The learners can order something.
  The learners can give/obtain information concerning place, time, quality, physical condition, availability, quantity and absence/presence.
- The learners can give personal information.
- The learners can ask for and offer assistance.

12
In the context of objectives 10-11 the learners can:
- greet somebody
- take leave of somebody

- introduce themselves and react to being introduced
- apologize and react to apologies
- thank somebody and react to being thanked
- congratulate somebody
- give somebody their best wishes
- invite somebody and react to invitations
- warn somebody
- accept something and decline/refuse something

13
In the context of objectives 10-11 the learners can:
- begin and end a conversation
- draw somebody's attention
- address somebody
- ask somebody to speak more slowly, ask for repetition
- signal non-understanding
- ask how something is said in the foreign language
- ask what something means
- begin and end a telephone conversation
- spell their own name and address

Writing skills

14
The learners can fill in simple forms such as hotel registration forms, stating their name, sex, dates of birth, arrival, etc., nationality, address, etc.

15
The learners can write
- a short formal letter requesting reservation or tourist information
- a simple personal letter
  - to arrange or cancel a visit or other contact;
  - to pass on a message;
  - to say thank you;
  - to keep in touch;
  - to give their best wishes.

In doing so learners meet the criteria given in objective 10.

B Compensatory strategies and techniques

16
The learners can avail themselves of various techniques and aids which may enable them to satisfy their communicative needs also in cases where their command of the language and/or of its use is insufficient.

Particularly:
- they can use compensatory interpretation strategies, e.g. in deducing the meaning of unknown elements from their context or on the basis of other characteristics;
- they can use compensatory production strategies, such as the use of paraphrase and other approximation devices;
- they have some reference skills, being able to use such aids as a word list, a dictionary, and a grammatical survey.

C Socio-cultural competence

17
The learners know in what countries or regions the foreign language is used as an important medium of communication, and have some insight into roles and importance of the foreign language in international contacts, both in the fields of commerce and technology and in those of social and cultural life, in their own country as well as elsewhere.

D Orientation in foreign language learning

18
The learners have some insight into typical features of life in those countries where the foreign language is the vernacular, and have been introduced to some typical verbal manifestations of (the) culture(s) of the areas concerned.

19
The learners can relate the learning of foreign languages to their own needs and experiences. Particularly:
- they have some insight into the objectives set to them and/or into those they might choose for themselves;
- they have some familiarity with procedures and aids which will contribute to the achievement of these objectives;
- they are capable of assessing, in ways appropriate to their own ends, their progress in foreign language acquisition.

Section II

Methodological Aspects of Language Testing

CRITERION-REFERENCED LANGUAGE TEST DEVELOPMENT: A PROLEGOMENON

Fred Davidson, University of Illinois at Urbana-Champaign
Brian Lynch, University of California, Los Angeles

The article examines and illustrates the application of criterion-referenced measurement (CRM) principles to test design. The writers present a formula for workshops in CRM test development that trains language teachers in writing criterion referenced tests. Extensive examples of actual test specifications are included.

1. PART ONE: THE CASE FOR ITERATIVE CRLTD

This paper has two goals. First, it argues that since the previous research in language testing related to criterion-referenced measurement (CRM) has focused on issues of test and item analysis, it would be productive to explore the application of CRM principles and techniques to the issues of test design and development. Second, it sketches a workshop formula for training teachers to do criterion-referenced language test development (CRLTD) in an iterative or cyclically improving fashion; it was this procedure that was followed at the Finland workshop.

Recent writing in language assessment has seen growing interest in criterion-referencing as an alternate epistemology to norm-referencing. Discussions of criterion-referencing appear in many scholarly papers and books on language testing, for example Hudson and Lynch (1984), Hudson (1989), Brown (1989), Hughes (1989), Bachman (1990), and Cook (1992). These writings discuss quite well the differences between criterion-referenced measurement (CRM) and the older, more pervasive form of assessment, norm-referenced measurement (NRM). We shall summarize the salient differences here.

A criterion-referenced test is one in which each item or task is securely tied to a specific student behavior, called a criterion, for which success/failure can be determined. In contrast, traditional tests are often called norm-referenced, in which success/failure is often determined by an examinee's ranking in a normal population of test takers. Logically, all human assessment has need for both types of tests. For example, consider a situation in which you have limited number of scholarships to distribute each year within an educational setting. From year to year, the quality of the students applying for these scholarships may vary; however, each year the committee responsible for awarding them

wants to know who the top five applicants are. The decision is relative to the applicant pool that exists that particular year, not to an absolute criterion of academic potential or achievement. This is logically a situation that calls for a ranking judgment: NRM. On the other hand, a good example of a criterion-referenced test is a driving exam; the examiner asks the driver to drive the car, and if she or he can do so (perform the behaviors – an absolute, not relative judgment), then the license is given. It would not be logical to state that the examinee ranked 85th among all drivers taking the exam. What is at issue is whether or not the examinee can drive. The focus of the measurement is a criterion, and calls for a CRM judgment.

Criterion-referenced tests can offer many advantages in test design (and can sometimes be integrated with norm-referenced qualities). We submit that the greatest advantage to criterion-referenced tests is that they state the expected behavior very clearly. This is done, most typically, through a very clearly written test specification or plan. A rubric for CRM specification has been developed by Popham (1978, 1981:Ch 9), and it is this rubric that guides section 2 of this paper below.

The effect of heightened focus on the expected behavior should be an increase in content validity, or the 'truth' of a test against a specified domain of content. The domain of content, we contend, can be either a proficiency content domain (c.f. the GTELP test utilized by Hudson, 1989) or achievement content domain by specifying to a course of study in a language teaching setting (cf. Hudson & Lynch, 1984). Furthermore, we suggest that research has shown that in language testing the statistical differences between CRM and NRM statistics are minimal, and that CRM and NRM analyses can be fused to great advantage. For example, Brown (1989) illustrates the development of a placement exam where items were selected based on both NRM and CRM criteria. The goal of his study was to

> ...combine the useful qualities of NRTs (Norm-referenced tests) with those of CRTs (Criterion-referenced tests) to create placement tests that not only spread students out along a continuum of language abilities (NRT) but do so on the basis of items that are demonstrably related to what the students learn when in the program. (Brown, 1989:73)

In his discussion of the results, Brown argues that he has demonstrated that a shorter test which fuses NRM and CRM qualities is a better measure of the construct validity of his placement setting as well as heightened validity to the content of instruction. This finding was echoed by Cook (1992) in further analyses from the same institution, the University of Hawaii. Based on his data, Cook reported that analytical differences of CRM and NRM were not helpful to advances in our understanding of language testing.

Brown (1989) goes on to state that statistical fusion of NRM and CRM is 'just a beginning' (p.79). In particular he notes:

> We must continue to strengthen this relationship (between testing and teaching) and attempt to understand it better. It may prove particularly fruitful to examine the curricular implications of our findings with a focus on what can be borrowed from the course objectives to refine further the relationship between the placement test and the course.
> (Brown, 1989:79)

In Part 2 of our paper, we offer a means to do precisely that. We suggest an iterative, bottom-up, consensus-based method which uses a version of Popham's specification framework in order to better reflect actual teaching objectives or proficiency guidelines.

2. PART TWO: CRLTD WORKSHOP DESIGN

Typical test specifications/plans under NRM or under-realized CRM might look like the following example for a hypothetical reading test:

Figure 1. Typical table of test specifications: A Reading Test.

Skill tested	Number of tasks / items per skill	Type of task / item per skill	Special Materials Comments
Vocabulary in context	20 items	multiple-choice with 2-3 sentence stem	No passage needed.
Scanning	20 items	multiple-choice: specific information from passage	Passages needed; 2-3 items per passage.
Skimming	20 items	multiple-choice: specific information from passage	Different passages from scanning. 1-2 items per passage.

Strictly speaking, the only guidance for the skill to be tested is the first column, labeled 'Skill Tested', though the third column also provides some insight into what is intended. Item- writers using this table of specifications would be flying

blindly. Using a hit-and-miss process, they may or may not produce comparable items.

We suggest that a much better characterization of the skill to be tested is provided by the following CRM specification (spec), here presented in its general form. We note that this rubric applies to each identifiable and unique skill. The GD, SI, PA, RA, and SS components follow the descriptions in Popham (1978, 1981: Ch.9). The Specification Number, Title, and Related Specification components relate to our experience with CRM specs in various workshops and in the development of the revised UCLA ESLPE and the GTELP exam (see Hudson, 1989). We have adopted the term 'PA' (prompt attribute) as used in Brown, Detmar, & Hudson, (1992) to avoid any confusion with 'stimulus-response' as part of behavioristic psychology.

Figure 2. Ingredients of a CRM Spec

- Specification Number: provide a short index number.

- Title of Specification: a short title should be given that generally characterizes each spec. The title is a good way to outline skills across several specifications.

- Related Specification(s), if any: list the numbers and/or titles of specs related to this one, if any. For example, in a reading test, separate detailed specifications would be given for the passage and for each item.

(1) General Description (GD): a brief general statement of the behavior to be tested. The GD is very similar to the core of a learning objective. The purpose of testing this skill may also be stated in the GD. The wording of this does not need to follow strict instructional objective guidelines; compare the two examples given below.

(2) Sample Item (SI): an illustrative item or task that reflects the specification, that is, the sort of item or task this spec should generate.

(3) Prompt Attributes (PA): a complete and detailed description of what the student will encounter. If the item is the selected- response type (see 'RA' below) then the PA must contain a complete and detailed description of each choice. Note: choices can be randomized later in test development.

(4) Response attributes (RA): a complete and detailed description of the way in which the student will provide the answer; that is, a complete and detailed description of what the student will do in response to the prompt and what will constitute a failure or success. There are two types of RAs:

a. selected response, e.g. a multiple choice item in which the student selects the best response from among several options.
b. constructed response: clear and detailed description of the type of response the student will perform.

(5) Specification Supplement (SS): a detailed explanation of any additional information needed to construct items for a given spec. In grammar tests, for example, it is often necessary to specify the precise grammar forms tested. In a vocabulary specification, a list of testable words might be given. A reading spec might list in its supplement the textbooks from which reading test passages may be drawn.

A well-written CRM spec, if given to a group of similarly trained personnel working in a similar setting, should produce a set of content-similar items or tasks. Each person, whom we call an 'item writer' should be able to produce an item or task that the 'spec writer' acknowledges as fitting the CRM spec. We next present two example CRM specs. We would note quite openly that these examples are not intended to be perfect or refined. Instead, they are offered as initial working drafts of specifications that will be critiqued in the workshop process that will be outlined in the following section. One way in which these specs are not perfect is that they may cause item writers to produce dissimilar items. Other flaws may also exist. We would now like to present a sample CRM spec taken from an operational test setting, the UCLA English as a Second Language Proficiency Examination (ESLPE) revision begun in 1988-89. The spec as presented in Figure 3 below is an early attempt at a global CRM description of the reading subtests (texts and items) for the revised ESLPE.

Figure 3. A Sample CRM Specification

Title: UCLA ESLPE TEST SPECIFICATION: Reading
Related Specs, if any: See UCLA ESLPE Spec. Bank.
Spec Number: 1-A (June 1989)

GD — General objectives to be tested: Students will demonstrate their ability to read academic general/non-technical reading text(s). They will be able to distinguish main ideas from supporting details, synthesize main ideas and specific details, and make inferences from assertions and arguments.

Specific Objectives to Be Tested: Students will demonstrate their ability to:

a. distinguish main ideas from supporting ideas;
b. recognize specific details which may be in the form of supporting statements, supplementary information, and numerical and/or graphical information;
c. synthesize information, which may be both main ideas as well as specific details, in order to understand assertions and arguments;
d. make inferences from statements and arguments, and numerical and/or graphical information.

SI There are two item samples below: multiple-choice (MC) and true-false (TF). Both sample items are based on "Pioneers! O Pioneers!" from "The Story of English" (see SS below).

(MC) The main idea of the first paragraph of the text is:

a. the English language has provided a way of understanding the Anglo-American rivalry
b. the English language has led to several battlegrounds between English and Americans
c. the Americans had to make up many new words as they were pioneers
d. the English and the Americans are different in every way of life

(TF) The Americans waited for the American Revolution to change their English.

PA Characteristics of the Reading Text: The reading text(s) for the test should have the following characteristics:

a. general, academic reading, ideally non-technical/not too specific;
b. authentic reading from, preferably, an introductory university text;
c. length: at least about 1500 words, maybe about as much as 2000 words, for each text, which can be divided into two or three sections;
d. language purpose: preferably, persuasive;
e. rhetorical mode: preferably, text(s) that use all of the following: descriptive, narrative and argumentative;
f. examples, illustrations, anecdotes, graphs, charts, descriptive statistics, figures, and tables (at least one or two of these would help the material approach "authenticity" if the text(s) are not taken from an introductory university text).
g. preferably, text(s) that have transparent organizational markers: section/division headings, chronological, sequential arrangements are examples.

Characteristics of the test items: The test items should have the following characteristics:

a. For Objective (a): distinguishing main ideas from supporting ideas. Main ideas may be characterized as main propositions or controlling ideas in the text and supporting ideas as particularizations, like illustrations, examples, anecdotes, and numerical or graphical information that work as detail. Suggestion: Main ideas

may be identified by several procedures; a few useful ways are: to read the material fast, to skim or scan the text(s), to read the first sentence of each paragraph, to look for section headings, to look for explicit organizational markers.

b. For Objective (b): recognizing specific details. Specific details should be supplementary information, supporting ideas, numerical and graphical information, examples, illustrations, tables, figures. Specific details should not be main propositions or controlling ideas. Caution: do not include the most trivial or abstruse piece of information in the text(s) that could very well be almost unimportant for the understanding of the text(s).

c. For Objective (c): synthesizing information. Synthesis of information should be a synthesis of one main idea with another idea, one main idea with supplementary ideas or one supplementary idea with one or other supplementary ideas. Caution: do not include difficult or complex conceptual synthesis or tedious mathematical computations.

d. For Objective (d): making inferences. Making inferences should involve the ability of students to draw meanings, implications, intentions, and possibly, opinions and attitudes of the author/text(s). Caution: do not include inferences that may require cultural, social or political knowledge or a particular world view or any other outside knowledge in order to make inferences.

Description of Item Format for the Test Items

a. For Objective (a): distinguishing main ideas from supporting ideas. This objective can be best tested with the multiple-choice (MC) format. Main ideas of the text should be paraphrased from the text, maybe from individual paragraphs or from sections, and presented as one of the three or four alternatives. Incorrect statements that make up the wrong choices could include 1) supporting ideas, 2) misstatements of main ideas, and possibly, 3) overgeneralizations. Caution: for main idea statements, do not "complexify" the main ideas in the paraphrase process and do not, on the other hand, provide "quotes" from the text(s). In addition, do not change the intent, opinion, implication, or meaning of the idea to be tested.

b. For Objective (b): recognizing specific details. This objective can be arguably best tested by the true-false format. Specific details chosen should be paraphrased (and then listed in the order of their appearance) for the item to work as "true" statements. For "false" statements, specific details of the text(s) should be paraphrased with

the details changed or added, or a misreading of information should be made. Caution: for True statements, keep the specific information to be tested "intact" despite the paraphrasing. Students will do truth assessment of the statements presented to them and indicate whether the statements are true or false in terms of the text read.

c. For Objective (c): synthesizing information. This objective can be tested with the MC format. The synthesized information to be tested should be paraphrased and presented in the form of a question, (and not in the form of a statement). Correct synthesis of information should be made by putting together pieces of information available in the text. Incorrect synthesis should be given in the three or four alternatives where (1) incorrect synthesis of information, (2) over-synthesis, and (3) synthesis of information not stated in the material are made.

d. For Objective (d): making inferences. This objective can be tested with the MC format. The inferential statements to be tested for their "correctness" should be directly inferable from the material. A few of the possible inferential types of questions are meanings, intentions, opinions, and attitudes. Inferences should be presented in question form and not in statement form. Incorrect inference types that could be included in the alternatives are: (1) misinterpretation of words, phrases, or expressions; (2) inference made from outside the text; (3) inference based on incorrect interpretation of detail; and (4) unrelated/irrelevant inference.

RA To successfully complete the criterion implied by each item, the student will select the correct choice from among the alternatives given (in an MC item) or implied (in a TF item).

SS For the text of the passage to which the sample MC item refers, "O Pioneers" - The Story of English, by McCrum, R.; Cran, W. & MacNeil, R. 1986 (pp.235-240) . NY:Elisabeth Sifeton Books/Viking.

****

Developed by Sara Cushing, Antony Kunnan, Charlene Polio, and Brian Lynch, June 1989.

Several refinements to this spec are possible, and in fact during the normal test development process several refinements evolved at UCLA. A revised version of this spec is given in Figure 4 on the next page.

Figure 4. Revised CRM Spec.

Title: UCLA ESLPE TEST SPECIFICATION: Reading
Related Specs, if any: See UCLA ESLPE Spec. Bank.
Spec Number: 1-A (March 1992)

GD — General Description of the Criterion Being Tested: Examinees will demonstrate their ability to read genuine academic texts from general, non-technical university level materials or from genuine non-academic texts with content and language similar in difficulty and sophistication to the academic texts.

The specific reading skills on which examinees will demonstrate their ability are:

a)recognition of main ideas (and distinguishing them from supporting ideas)
b)synthesis of information across more than one paragraph in the text
c)recognition and recovery of information in the form of specific details
d)recognition of opinions (and distinguishing them from information presented as fact)
e)recognition of inferences drawn from statements and information presented in the text
f)identification of the meaning of key vocabulary items in the text.

PA — Requirements For the Text: The reading texts for the test should have the following characteristics:

a) academic content from an authentic university level text that is not overly specific to a particular field and does not contain extremely technical vocabulary or, general interest content from an authentic text that is aimed at a college-educated audience
b) a self-contained section (with minimal or no editing to the original text) of 800 to 1200 words in length
c) several rhetorical modes such as description, narration, argumentation, with a clear overall organizational structure (i.e., transparent markers such as section headings and logical connectors)
d) examples, illustrations, anecdotes, graphs, charts, figures, and/or tables.

Description of the Test Items. The examinees will be presented with written selected-response items of two types: true-false and multiple-choice. Additionally, a special MC item type called 'vocabulary-in-context' will be used.

True-False Items.
The True-False format will be used for testing recognition of specific information, definitions, and opinions (and distinguishing them from statements of fact), or inferences in the text. The purpose of having True-False items is two-fold: to present propositions for "truth assessment", attempting to simulate

the hypothesis testing that is an important part of the reading process, and to allow for an easier generation of items than multiple-choice (MC)-only format allows. The True-False items will consist of a complete sentence statement, followed by two options: A. True and B. False. True statements will be constructed from information in the text. False statements will be constructed by taking information from the text and changing some aspect of it or rewriting it to be the opposite of the original information.

MC Items.
The MC format will be used for recognition of main ideas (and distinguishing them from supporting ideas), synthesis of information across more than one paragraph in the text, as well as recognition of definitions and opinions (also covered by the True-False format). Each MC item will consist of a wh-question followed by four options, which may be either complete sentences or single phrases. One of the four options will represent the best answer to the question, according to the information in the text. The other three options (distracters) will depend on the type of question being asked.

- If the question is asking for a recognition of a main idea, distracters will be 1) statements which are clearly over-generalized or 2) ones which express ideas mentioned in the text, but are supporting ideas or otherwise not the main idea in question.
- If the question is asking for a synthesis of information, the distracters will be 1) statements which put information together that does not cohere for the text in question or 2) ones that represent a synthesis of information that goes beyond that presented in the text.
- If the question is asking for a recognition of definition or opinion, the distracters will be 1) statements that confuse statements of fact in the text with statements of opinion (or vice versa) or 2) ones that present definitions, facts, or opinions that do not occur in the text.

Vocabulary in context items.
The multiple-choice (MC) format will be used to test vocabulary. These items are considered vocabulary in context because the specific vocabulary to be tested comes from the reading passages and, therefore, has a context. It is not assumed that the examinees will be guessing from context in order to answer the questions. It is assumed that the vocabulary selected for testing will be unfamiliar to many, if not most, of the examinees and that they will use various strategies including clues from the reading context in order to identify the correct meaning of the test item. Vocabulary items chosen for testing should therefore be: a) unusual or low-frequency words which non-native speakers would not be likely to have come across in other reading b) embedded in the context of the reading passage (e.g., a word that is being defined; a word that has an accompanying example or illustration; a word that is being compared or contrasted to another, more familiar word or phrase. Vocabulary items selected for testing will be underlined in the reading passage (for which the paragraphs will be numbered). The test item will have the following prompt: The word "_____" in paragraph ___ is closest in meaning to:

a) _____
b) _____

c) _____
d) _____

The options ("a" through "c") should all be words which the examinees can be expected to know outside of the reading passage context because of their frequency and simplicity. All options should be from the same word class as the selected vocabulary item. In addition to the correct option, there should be three distracters drawn from the following types:

- a different sense of the selected vocabulary item which is inappropriate to the context
- a semantically unrelated word which is phonologically or morphologically similar to the selected vocabulary item
- a word that makes sense in the immediate context (that is, within the particular phrase or sentence in which the selected vocabulary item is located in the reading passage) but does not make sense within the intended, larger context of the reading passage

RA Description of the Answer Format The examinees will have a separate answer sheet/card on which they will darken in their selection (A or B; A,B,C, or D) with a #2 lead pencil (for machine-scoring). They will be instructed to indicate their answer by marking the one best alternative for each test item.

SI There are two item type samples below: multiple-choice (MC) and true-false (TF). TF items are from Miller, M.M. "Man, the Bridge Builder" in Introduction to Engineering, Glorioso/Hill 1975. (pp 33-37) Prentice-Hall in Academic & Study Skills for International Students, Rosenthal, L & Rowland, S.B. 1986. Englewood Cliffs,NJ: Prentice-Hall. MC items are from "Cholesterol" by Isaac Asimov, LA Times, 1986, in the St. Martin's Guide to Writing 1988 (pp.152-153) Short Second Ed. NY:St. Martin's Press.

(TF) The author implies that early man was more patient than modern man.
A. True
B. False

(TF) The first attempt at the arch bridge was in about 4000 B.C.
A. True
B. False

(MC) Which paragraph best expresses the main purpose of this essay?
A. Paragraph 10
B. Paragraph 11
C. Paragraph 13
D. Paragraph 16

(MC) What is an axon?
A. A kind of nerve cell
B. A part of a nerve cell
C. A by-product of nerve cells
D. An insulator for nerve cells

SS Additional details:

Re. requirements for the reading text:

- Each Reading & Vocabulary subtest should have two reading passages: one reading should have content from the natural sciences and the other should have content from the social sciences.

Re. description of test items:

- Specific information is supplementary information, supporting ideas, examples, illustrations, numerical information, and information retrievable from tables and figures. For the purposes of the test, it should not include the most trivial or obscure piece of information in the text, especially if it is unimportant for the understanding of the text.
- Inferences refer to meanings, implications, intentions, and, possibly, opinions and attitudes of the author that are not explicitly stated but are recoverable from the text. For the purposes of the test, these should not include inferences that may require cultural, social or political knowledge beyond that assumed to be background knowledge possessed by the students taking this exam.
- Main ideas may be characterized as main propositions or controlling ideas in the text and supporting ideas as particularizations, like illustrations, examples, anecdotes, and numerical or graphical information that work as detail. Suggestion: identify potential main ideas by the following procedures: read the material fast, skim or scan the text(s), read the first sentence of each paragraph, look for section headings, look for explicit organizational markers.
- Synthesis of information is a linking of one main idea with another idea, one main idea with supporting ideas or one supporting idea with one or more supporting ideas. For the purposes of the test, this should not include difficult or complex conceptual synthesis or tedious mathematical computations.

This specification was revised in March, 1992 by Brian Lynch. Previous versions of the specification were developed by Sara Cushing, Antony Kunnan, Charlene Polio, and Brian Lynch.

We would like to highlight a few of the characteristics that evolved in this spec between 1989 and 1992. First, it is interesting to note that the UCLA group adapted the general CRM Spec. ingredients (see Figure 4) to suit their own needs; for example, the GD is split into discussion of general and specific skills. Furthermore, level of detail represented by this spec changed between 1989 and 1992. This change reflected the perceived need to make such references as "main idea" and "synthesizing information" clearer than they had been in the 1989 version. In addition, the number of reading skills specified in the GD increased. Similarly, the number of sample items increased, with more detail specified in the format for the True-False items, reflecting the need of item writers for greater item-writing clarity. Additionally, the content of the spec changed. The clearest example of content change is the addition of a new subskill: vocabulary-in-context. The instructional personnel felt it necessary to include and describe that as a key component to the reading assessment. Another interesting change is the reading texts that the sample items are drawn from. By

1992, the content came from more scientific academic disciplines (engineering and bio-chemistry, although the latter is not terribly technical in its discourse). Between 1989 and 1992 the instructional personnel had decided upon different genres of academic reading.

We contend that no spec is ever finished. The 1989/1992 ESLPE reading spec may continue to evolve. Perhaps greater refinement of the contextualized vocabulary skill is possible. Perhaps, also, the spec is becoming unwieldy and is evolving into several specs; for example, perhaps the personnel might decide to split the spec into several new specs as indicated by the specific skills listed in the GD. This spec has evolved and is evolving through a bottom-up evolutionary CRLTD process.

We suggest that instructional personnel at a language teaching setting gather together to hold a test development workshop. The workshop will serve to emulate the group discussions that evolved the 1989/1990 ESLPE spec above. We call this 'iterative CRLTD', and suggest that it be organized loosely along these two guidelines:

(1) What are the skills we wish to test? What is our 'mandate'? Mandate refers to the curriculum or proficiency guidelines that direct the instruction at that setting. Often this mandate is set down in documentation. For example, the ACTFL (ACTFL, 1986) guidelines act as a mandate for a particular approach to curriculum design, instruction, and testing in certain modern language teaching contexts. Sometimes, the CRLTD process will evolve further refinements of the mandate. For example, it might be discovered that the types of language tasks suggested by the ACTFL guidelines are inappropriate for a particular instructional context. The CRLTD process could function to develop descriptions of criteria to be tested that relate more closely to the language proficiency of the students in that context and their ultimate needs and goals for the language. Finally, the CRLTD process will sometimes result in an articulation of a mandate where one did not previously exist in a formal sense. Each skill, stated as an instructional objective, becomes the General Description of a CRM spec.

(2) How can we test each skill? This is best addressed by selecting a few sample skills and attempting to write a CRM spec from each. Instructional personnel in the institution should form small workgroups. Each workgroup should write one spec. The groups should then exchange specs and each group attempt to produce an item or task from another group's spec. The groups should reassemble as a whole and discuss needed changes. Figure 5 illustrates this process.

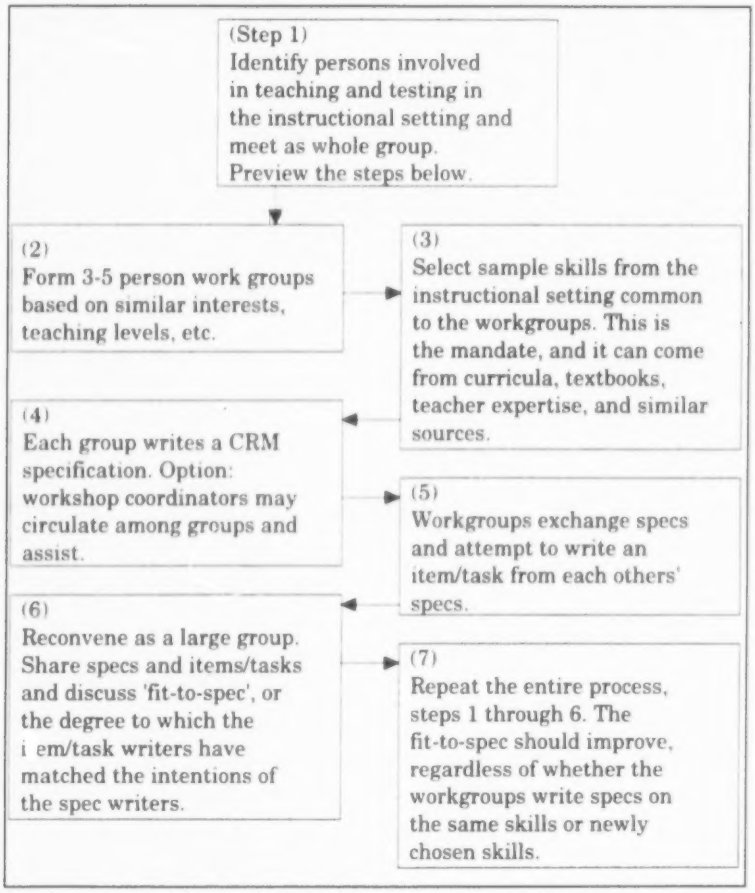

(Step 1)
Identify persons involved in teaching and testing in the instructional setting and meet as whole group. Preview the steps below.
(2)
Form 3-5 person work groups based on similar interests, teaching levels, etc.
(3)
Select sample skills from the instructional setting common to the workgroups. This is the mandate, and it can come from curricula, textbooks, teacher expertise, and similar sources.
(4)
Each group writes a CRM specification. Option: workshop coordinators may circulate among groups and assist.
(5)
Workgroups exchange specs and attempt to write an item/task from each others' specs.
(6)
Reconvene as a large group. Share specs and items/tasks and discuss 'fit-to-spec', or the degree to which the i em/task writers have matched the intentions of the spec writers.
(7)
Repeat the entire process, steps 1 through 6. The fit-to-spec should improve, regardless of whether the workgroups write specs on the same skills or newly chosen skills.

Figure 5. Steps in a CRLTD Workshop.

Most importantly, as noted in Step (7) in Figure 5, this entire cycle should be repeated several times. Ideally, any given spec should be written, tried by another group, returned to the spec writers for revision, then tried out by item writers again. Ideally, too, that whole process should be repeated for a new spec. By the end of two or three complete cycles, all participants will be communicating more freely and fostering a better climate of bottom-up language test development.

Are there any predictable difficulties with iterative CRLTD? What dilemmas might turn up during such a workshop? Our experience has suggested that three common problems tend to arise. The following represent suggested solutions to those problems.

(a) Write the spec at the correct level of generality. Don't write an *item* instead of a spec, and conversely, don't write a spec that is so general that it offers item writers too little guidance.

(b) Don't confuse the PA and RA, that is, keep clear the distinction between what the examinee will encounter (the PA) and what the examinee will do as a result of that (the RA).

(c) Constantly, while writing a spec, consider the nature of the potential item writer. This implies questions such as the following: Does the item writer have access to the necessary materials to produce an item from the spec? Does the item writer have adequate training for the spec? Is the spec *clearly written* — does it state what it wants in unambiguous language?

3. PART THREE. CONCLUSION: ITERATIVE CRLTD, GROUP DYNAMICS AND PROACTIVE TEST REFORM

Cyclical application of CRLTD in the fashion described here has some interesting benefits common to all group-based language test development. We do not claim that iterative test development is a new idea. We suggest, however, that careful application of CRM specifications during test development does promote refreshing, interesting and positive changes in language education settings. We base this contention on many hours of application of this system in settings ranging from 1 or 2-hour introductory workshops (e.g., the Finland conference on which this volume is based), to longer 2-3 day workshops (in Egypt and Guatemala) or even to operational language test design (as with the current UCLA ESL Placement Exam).

There are several specific ways in which iterative CRLTD can make a positive contribution to both testing and curriculum development. First, iterative CRLTD proactively fosters cooperative group dynamics. It gives a voice to more people than may normally have a voice in the testing process. While we see this as a benefit, we openly acknowledge that highly administrated, top-down test development settings may find a certain amount of difficulty in attempting to apply the principles we suggest here. In contrast, it has worked successfully in the UCLA ESL Service Course context. There, the current UCLA ESL placement exam continues to be developed following the CRLTD process, which involves Teaching Assistants, non-TA graduate students interested in language testing, TA supervisors and the ESL Director. The test development policy in the UCLA ESL service courses now incorporates ongoing, bottom-up CRLTD precisely in order to encourage wider input to testing by all instructional personnel.

Second, iterative CRLTD is reactive. It allows changes in language teaching and acquisition theory to find a voice more quickly than might otherwise occur. Personnel who have participated in CRLTD workshops have been able to coordinate and evolve newer, task-based CR tests that may better reflect

changes in the modern communicative language teaching era. In a sense, this is a type of washback, or backwash, that has the opposite direction of that usually discussed in the literature (e.g. Hughes, 1989:ch. 6). Here, the curriculum or instructional context can act to change the test, rather than vice versa. Put more simply, CRLTD fosters content validity. In fact, the UCLA context is an example of this type of curriculum to test washback. The current form of the test, and the use of CRLTD, was the result of requests by TAs for a placement exam that reflected the types of tasks and texts that were used in the ESL curriculum (which, in turn, reflected the types of tasks and texts that were used in the academic courses at UCLA).

In conclusion, iterative CRLTD is a means to help with the most difficult aspect of assessment: the leap from theory to operationalized measurable task, which, we contend, is a very central issue to good testing (see Davidson, Hudson and Lynch, 1985). In this paper we offer iterative CRLTD as a simple work-shop-like process to all who wish to try it, and we encourage all who do so to contact us with further comments on its success or failure.

Acknowledgements:

This paper is a writeup of the philosophy and procedures which guided the Criterion-Referenced Test Construction workshop at the Symposium/Seminar entitled 'Language Testing in the 1990s', which was held at Jyväskylä, Finland, from 26 Aug through 31 Aug 1992. The authors gratefully acknowledge the audience commentary at that meeting.

REFERENCES

American Council on the Teaching of Foreign Languages (ACTFL). 1986. *ACTFL Proficiency Guidelines.* Hastings-on-Hudson, NY: ACTFL.

Bachman, L.F. 1990. *Fundamental considerations in language testing.* Oxford: Oxford University Press.

Hughes, A. 1989. *Testing for language teachers.* Cambridge: Cambridge University Press.

Brown, J.D. 1989. Improving ESL placement tests using two perspectives. *TESOL Quarterly* 23:1, MAR 1989, pp.65-84.

Brown, J.D., E. Detmer, and T. Hudson. 1992. *Developing and validating tests of cross-cultural pragmatics.* Paper presented at the Fourteenth Annual Language Testing Research Colloquium, 27 FEB - 01 MAR 92, University of British Columbia, Vancouver, British Columbia, Canada.

Cook, G. 1992. *The place of placement tests.* Paper presented at the 26th Annual Convention of Teachers of English to Speakers of Other Languages (TESOL), 03-07 MAR 92, Vancouver, British Columbia, Canada.

Davidson, F., T.Hudson and B. Lynch. 1985. Language testing: operationalization in classroom measurement and L2 research. In Marianne Celce-Murcia (ed.) *Beyond Basics: Issues and Research in TESOL.* Rowley, MA: Newbury House. pp. 137- 152.

Hudson, T. 1989. *Measurement approaches in the development of functional ability level language tests: Norm-referenced, criterion-referenced, and item response theory decisions.* Unpublished Ph.D. Dissertation, University of California, Los Angeles.

Hudson, T. and B. Lynch. 1984. A criterion-referenced measurement approach to ESL achievement testing. *Language Testing* 1:2, December 1984, pp. 171-201.

Popham, W.J. 1978. *Criterion-referenced measurement.* Englewood Cliffs, NJ: Prentice-Hall.

Popham, W.J. 1981. *Modern educational measurement.* Englewood Cliffs, NJ: Prentice-Hall.

A METHOD FOR IMPROVING TASKS ON PERFORMANCE-BASED ASSESSMENTS THROUGH FIELD TESTING

Dorry Mann Kenyon and Charles W. Stansfield
Center for Applied Linguistics
Washington, D.C.

In this article on test development, we present a method for systematically collecting information from examinees and raters to improve tasks used in performance-based assessments of language proficiency. Field testing standardized tests and using item-level information (for example, indices of difficulty and discrimination) for test revision is a long accepted practice in the development of multiple-choice tests. Such indices are limited, however, in informing revisions for individual tasks on a performance-based test. This article describes an efficient method of collecting quantitative and qualitative information about test tasks from both examinees and raters. The method is illustrated through examples from the development of a performance-based assessment of second language speaking proficiency. The benefits of using this method are discussed, particularly in situations where high-stakes performance-based tests are under development.

INTRODUCTION

In this article, we present a systematic method for assessing the quality of tasks on performance-based assessments of language proficiency and for improving them through collecting data from both examinees and raters. We define a performance-based language assessment as one in which an examinee is required to demonstrate language ability in a holistic manner by producing language in response to some means to elicit the performance. The means to elicit performance is what we call the task. The task may be a situational role-playing card, a written prompt or question, or a simulation task. The goal of the task is to elicit (or help in eliciting) the desired linguistic performance from the examinee. The performance-based assessment may consist of a single task, such as in the direct assessment of writing using an essay test with a single question, or it may consist of multiple tasks, as in the second language speaking test discussed here.

For more than a decade, interest in and support for the use of performance tasks in standardized testing has grown steadily among the public and the measurement community. Direct writing assessments, for example, are now common in most states in the U.S. Many occupational licensure and certification programs are moving toward incorporating either work samples or simulations of real life professional tasks into their testing programs. In second language proficiency testing, the well-known Oral Proficiency Interview (OPI) of the Interagency Language Roundtable (ILR) and the American Council for the Teaching of Foreign Languages (ACTFL) makes extensive use of role-playing situations. In spite of the interest in performance-based assessment, there is still a paucity of published information on the development of quality tasks.

The development of tasks for a performance-based assessment involves a number of stages. These include determining the overall test specifications, writing the specifications for the tasks, training task writers, reviewing and revising both tasks and preliminary drafts of the test, selecting the criteria for scoring, developing scoring scales, and selecting and training raters. This article focuses on one aspect of the development of a performance-based assessment: the collection of data from multiple sources during the field testing of tasks in order to inform decisions about discarding, retaining and revising tasks. This aspect of performance test development is largely ignored in the literature, and current practice often fails to capture useful and important information.

1. PROBLEM

On a performance-based assessment of language proficiency, particularly one that may have important repercussions for the examinee as an individual and society as a whole, the tasks used to elicit examinee performance must have two characteristics that are critical to validity and accuracy of measurement. First, in order to ensure that the test is fair and unbiased, the tasks must allow, to the fullest extent possible, each examinee a fair and equal opportunity to give the best possible demonstration of his or her ability. Second, in the interest of obtaining the most accurate measurement of ability possible, these tasks must elicit a performance sample that enables scorers to adequately evaluate the examinee. Test developers are thus faced with the problem of ensuring that the tasks on their performance-based measurements exhibit these characteristics. Evidence for these characteristics cannot be found in traditional indices of item difficulty and discrimination alone. The specification of clear guidelines for writing tasks, the thorough training of the writers, and the review of tasks by experts aid in the development of appropriate tasks. However, these steps alone cannot guarantee that the tasks exhibit these two critical characteristics. Just as field testing is necessary in the development of objective paper and pencil measures, field testing is also required for performance-based assessments as well. No systematic technique has been adopted by test developers to ensure that these goals are being reached through field testing. In fact, Stiggins (1987),

in his excellent step-by-step module on the design and development of performance-based assessments, does not mention field testing as part of the development process. This is also the case with other thorough treatments of performance assessments (e.g, Haynes and Wilson, 1979).

While some comprehensive treatments of performance assessment do mention field testing, the treatment is cursory at best. For example, in the well known anthology by Berk, only the chapter by Siegel (1986) mentions pretesting and revision, and only five lines are devoted to this. Priestley (1982), whose employer, National Evaluation Systems, has been involved in the development of many state performance assessments for teacher certification, devotes only two short paragraphs to pretesting. He indicates that it can be used to identify unexpected examinee behaviors, to test any equipment required, and to determine the reliability of the instrument. It is especially surprising today to see that few discussions of writing assessment mention field testing, and even those that do, do not surpass the cursory treatment given it by Coffman (1971) in his otherwise classic article. As Coffman portrays it, feedback of the task under consideration is obtained during field testing exclusively from raters. Our conversations at AERA and NCME conferences with professionals involved in performance assessment indicate that examinee feedback is not sought and there is skepticism as to whether the examinee can actually contribute anything useful to the design of such instruments.

In this article on test development, we propose a method of obtaining quantitative and qualitative data during the field testing of a performance-based assessment that can either aid in or confirm the development of high quality performance tasks. Additionally, this method provides for the involvement in the test development process of both examinees and raters; i.e., those who play the most direct role in the operational assessment.

To illustrate this method, we have provided examples from the development of parallel forms in both French and Spanish of the Texas Oral Proficiency Test (TOPT), a performance-based assessment of spoken language proficiency required for the certification of Spanish, French, and bilingual education teachers in Texas. The TOPT was developed by second language test development specialists at the Center for Applied Linguistics (CAL) under a contract with the Texas Education Agency. It is one of eight tape-mediated, simulated oral proficiency interviews (Stansfield, 1989) that CAL has developed for the assessment of speaking ability in different languages under federal grants and state contracts since 1986. Each of the four forms of the TOPT used in the field testing stage elicited examinee speech performance on 16 tasks via a master tape and a test booklet. The first task involved ten relatively easy, short-answer questions linked to a single speaking context. They are intended as a warm-up exercise for the examinee and are not scored. The remaining 15 tasks function as prompts to elicit multiple samples of examinee speech. Each focuses on a different language function or speaking task. For example, in one task examinees are asked to give directions on how to get from one place to

another. In other tasks, examinees must give advice or present a factual summary about a given topic. These 15 tasks are divided into three groups of five: Pictures, Topics, and Situations. Examinee responses to all tasks are recorded on a second tape (the examinee response tape). That tape is scored at a later session by trained raters using the *Guidelines for Speaking Proficiency* of the American Council on the Teaching of Foreign Languages (ACTFL, 1986). The *Guidelines* describe a holistic scale for rating oral proficiency in a second language, commonly known as the ACTFL Scale.

The TOPT field testing used to illustrate this method involved four forms of the two tests (French and Spanish), 160 examinees and eight raters. The pretest examinees were adults preparing to teach French, Spanish or bilingual education in the state of Texas. In a few cases, the pretest examinees were already certified teachers. Examinees were tested at eight different sites in Texas. The six Spanish raters and two French raters were all very familiar with the ACTFL Scale and were instructed in the purposes of the field test rating. In addition to awarding a single overall rating to each examinee on the basis of his or her performance profile, the raters evaluated each individual task. The utility of their evaluation in the field testing process is part of the focus of this article.

2. METHOD

In the approach to field testing performance tasks we employ, quantitative and qualitative data are collected from both examinees and raters during field testing. Once items or tasks for the performance-based assessment have been developed, reviewed, and revised, the test form(s) is assembled. The preliminary version(s) of the performance-based test is then field tested and scored under conditions reflective of the operational program. To obtain additional information relevant to the improvement of the test, two supplemental data collection instruments are developed and administered during the field testing: one is completed by the field test examinees and the other by the field test raters. Each of these two instruments consists of two parts. The first part gathers quantitative data and is formatted on a machine-readable response form. The second part collects qualitative data and allows examinees or raters to write comments. Each of these instruments is discussed in turn.

2.1. Examinee Feedback Forms

A two-part instrument for the field test examinees is used to detect anything in the test (as a whole and for each individual task) which may be preventing examinees from giving their best possible performance. It must be designed with the age and maturity of the test takers in mind. Instruments designed for

children may be quite different from those designed for young adults, for example.

2.1.1. Quantitative Data Collection

The quantitative part of the feedback form is similar to an attitude questionnaire. It presents examinees with a series of statements about issues salient to the test and each of its tasks. After taking the performance-based test, examinees respond to these statements, indicating the degree of their agreement or disagreement with each statement on a five point scale. A "5" reflects the most positive response to the statement and a "1" reflects the least positive.

An example from the TOPT will illustrate the quantitative section of this feedback instrument. For the TOPT, the two most salient issues for every task were 1) whether examinees felt that the task allowed them to produce speech indicative of their current level of linguistic ability; and 2) whether they felt that the time allowed for them to prepare their answer and give their response was adequate. These concerns were presented to the test-takers in the following two statements: "As a whole, I felt **picture task #1** allowed me to give a response that reflects my current ability to speak the language," and "In **picture task #1**, the time allowed for preparing my answer and making my response was appropriate." These two statements were repeated for each task on the test, substituting the appropriate task designation for "picture task #1."

Five of the TOPT tasks involved the use of pictures. Statements about the clarity and interpretability of these pictures were developed for this instrument as well. For example, picture task #1 involved the use of a map of a section of a town, based on a bird's-eye view. In addition to the two statements given above, each examinee responded to a third statement on picture task #1: "The map for **picture task #1** was clear and understandable." This statement was repeated, with the appropriate task designation, for each of the five tasks involving pictures.

Statements about salient aspects of the test in general may also be included in the questionnaire. In the case of the TOPT, these statements addressed the issues of examinee nervousness, the use of a target language prompt (i.e., a question or statement as a signal to begin one's answer), and the degree to which the examinee is satisfied that a rater listening to the examinee response tape would get an accurate picture of his or her ability to speak French or Spanish.

For the TOPT, the entire machine-readable questionnaire contained 40 statements. The General Purpose Data Sheet II (Form 70921) from National Computer Systems (NCS) was run through a photocopier and overprinted with the question number and as much of the statement as would fit on one line (see Appendix for the complete sheet). A separate booklet contained the complete statements, as well as all instructions for completing the form and some

questions soliciting pertinent background information. The answers to these questions were also recorded on the machine-readable data sheet.

Calculating mean examinee responses to these statements provides quantitative data to assess the quality of the tasks from the examinees' point of view. This data then informs the task and test revision process. As an example, Table 1 below gives all the mean examinee responses for the tasks presented on Form D of the Spanish TOPT. This table shows that examinees responded differentially to the statements about the tasks. On this form, the field test examinees awarded their highest rating (mean = 4.22) to the statement about the clarity of the picture used for picture task #4, while the lowest rating (mean = 3.19) was awarded for the time allotted to answer picture task #2.

In revising the TOPT, special note was taken of all statements receiving a mean rating of 3.75 or less. These identified for the test developers tasks requiring careful review. Table 1 shows that for Form D of the field test version of the Spanish TOPT, 14 of the questionnaire items received a rating below 3.75, involving 8 of the 16 tasks. Of these items, one was about the quality of the picture used (for Picture Task 5), five were about the general adequacy of the task, and eight were about the amount of time allotted to respond to the task. Table 1 next page shows that for Picture Task 5, all three aspects of the task (general adequacy, picture quality and time) received ratings below 3.75, identifying for the test developers a task that definitely needed careful review and revision. Similarly, both statements (general adequacy and time) for Topic Task 3, Topic Task 4, and Situation Task 3 received ratings below 3.75, again indicating tasks needing careful scrutiny in the revision process. Thus, as can be seen, the quantitative part of the examinee feedback form allows the test developer to obtain an index of the degree to which pretest examinees rate each performance task as a fair and appropriate measure of their overall ability.

2.1.2. Qualitative Data Collection

The second part of this instrument provides examinees with a form containing spaces for written comments. The form asks examinees to explain why they awarded any item a lower rating (3, 2, or 1) in the quantitative part of the questionnaire and to suggest revisions that would improve the task. Examinees are also encouraged to comment on general statements about the test and to write about any additional concerns not addressed elsewhere.

Table 1. Example of Field Test Examinees' Mean Ratings on Quantitative Questionnaire (from TOPT-Spanish Form D)

Item	Task-Aspect	N*	Mean
ITEM1	Opening Conversation--General	27	3.963
ITEM2	Opening Conversation--Time	26	4.077
ITEM3	Picture #1--General	27	4.000
ITEM4	Picture #1--Picture	27	4.037
ITEM5	Picture #1--Time	26	4.038
ITEM6	Picture #2--General	26	3.423•
ITEM7	Picture #2--Picture	27	3.815
ITEM8	Picture #2--Time	27	3.185•
ITEM9	Picture #3--General	26	3.885
ITEM10	Picture #3--Picture	27	4.185
ITEM11	Picture #3--Time	27	3.852
ITEM12	Picture #4--General	26	4.115
ITEM13	Picture #4--Picture	27	4.222
ITEM14	Picture #4--Time	27	3.778
ITEM15	Picture #5--General	26	3.692•
ITEM16	Picture #5--Picture	27	3.593•
ITEM17	Picture #5--Time	27	3.519•
ITEM18	Topic #1--General	25	3.720
ITEM19	Topic #1--Time	27	3.815
ITEM20	Topic #2--General	27	3.889
ITEM21	Topic #2--Time	27	3.778
ITEM22	Topic #3--General	26	3.577•
ITEM23	Topic #3--Time	27	3.630•
ITEM24	Topic #4--General	26	3.577•
ITEM25	Topic #4--Time	26	3.577•
ITEM26	Topic #5--General	27	3.926
ITEM27	Topic #5--Time	27	3.704•
ITEM28	Situation #1--General	26	4.038
ITEM29	Situation #1--Time	26	3.846
ITEM30	Situation #2--General	27	3.778
ITEM31	Situation #2--Time	27	3.704•
ITEM32	Situation #3--General	26	3.654•
ITEM33	Situation #3--Time	26	3.577•
ITEM34	Situation #4--General	27	3.815
ITEM35	Situation #4--Time	27	3.556•
ITEM36	Situation #5--General	26	3.923
ITEM37	Situation #5--Time	26	3.769

• Questionnaire items receiving a mean rating below 3.75

* Not all of the 27 examinees who took Form D responded to every item on the questionnaire

During the field-testing of the TOPT, these written comments shed light on lower rated tasks and contained many useful suggestions for revising tasks that were seen as flawed. For analysis purposes, we organized these comments, typing them into a database coded with examinee, test form and task information. In addition, at the time of data entry, the person key entering the comments also entered a rating on a scale of 1 to 3 as to the degree of positiveness or negativeness the comment reflected. This rating was then used to order the comments from most positive to most negative. The ordered comments were then printed for use in the revision process. Often the examinees' comments drew attention to minor unclear details that were not caught by the test developers or any of the three external review committees. For example, in one picture sequence examinee comments pointed out that the story line was unnecessarily difficult to follow due to the similar appearance of two of the characters. To remedy this situation, the artist changed the clothing and appearance of the two characters.

2.2. Rater Feedback Forms

We design the second two-part instrument for raters of the field test performances. They use it to indicate the adequacy of each performance as actually elicited by each test task in terms of the information it provides for evaluating the examinee (using the designated scale or criteria).

2.2.1. Quantitative Data Collection

In the first part of the instrument, each field test rater is asked to indicate the adequacy of the performance on each task for assigning a score to each examinee rated. Some tasks may be generally more successful in eliciting a ratable performance sample than others. The mean rating of performance adequacy across examinees identifies strong and weak tasks from the perspective of the rater.

In the TOPT example, a rating of 1 represented a contribution that was "not useful" to the rater in assigning the examinee an overall score, 2 represented a "useful" contribution, and 3 a "very useful" contribution. For the 15 tasks on the Spanish Form D, for example, the mean of the ratings by the four field test raters ranged from a high of 2.85 for Picture #4 to a low of 2.55 for Situation #3. Low mean ratings identify tasks to that may be problematic and need revision.

The rater's instrument can also ask raters to provide quantitative information about other salient aspects of the test format. In many performance-based assessments, the time examinees have to complete a task may necessarily be limited. From the examinees' point of view, frustration may result if they feel they did not have enough time to complete a task. Likewise, raters may feel they need to have more of a performance sample from a task than has been allotted in order to make an adequate rating. In the case of the TOPT, it was

necessary to confirm the adequacy of the timed pauses since the amount of time allotted to a task can limit the examinee's ability to complete it. Thus, the raters were asked to provide quantitative information on whether the time limit for each task posed a problem for the examinee. If the examinee had more time than necessary to make a complete and adequate response to a task, the rater indicated the amount of extra time, in number of seconds, that the examinee had. If the examinee needed more time to demonstrate ability (or inability) to perform the language function presented in the task, the rater estimated the number of seconds needed. The data from the raters' form was then input into a database, after which tables were produced summarizing raters' evaluations of each item across all examinees. A frequency distribution of "seconds extra" (positive numbers) or "seconds needed" (negative numbers) revealed whether patterns of too long or too short time limits existed in the group of examinees field tested. This informed revisions to the time limit on each speaking task. Table 2 shows an example of these tables for two tasks. The data for Question #10 of the Opening Conversation indicates that the response time for that task should be lengthened. Only 50% of the examinees had an adequate amount of time. In contrast, the data for Picture Task #1 indicates that slightly under 50% had much more than enough time to give an adequate response for this task. About 36% had 5 or more seconds to spare.

Table 2. Example of Raters' Quantitative Time Data for Two Tasks Rated for 22 Examinees (from TOPT-Spanish Form A)

Task = Opening Conversation Question #10

Time	Frequency	Percent	Cum. Freq.	Cum. Percent
-15	1	4.5	1	4.5
-7.5	1	4.5	2	9.1
-5	6	27.3	8	36.4
-2.5	3	13.6	11	50.0
0	11	50.0	22	100.0

Task = Picture Item #1

Time	Frequency	Percent	Cum. Freq.	Cum. Percent
-7.5	1	4.5	1	4.5
0	11	50.0	12	54.5
5	2	9.1	14	63.6
7.5	3	13.6	17	77.3
10	1	4.5	18	81.8
12.5	4	18.2	22	100.0

Notes: Time is in seconds; negative amounts of time indicate raters' estimate of additional time needed for examinees to give a rateable response; positive amounts of time indicate raters' estimate of silence after examinees have given a rateable response.

2.2.2 *Qualitative Data Collection*

Qualitative data is collected by encouraging raters to make written comments on the second part of this instrument on problems in the performances of examinees. For example, they can mention if anything in the task seems to hinder the examinee from giving an appropriate performance, if the task elicits the desired type of performance, or if any type of examinee (male versus female, Hispanic versus non-Hispanic) seems to have an advantage or disadvantage in responding.

For analysis of the TOPT, the raters' written comments on the examinees' performances on the tasks were coded according to whether they were favorable, neutral, or unfavorable, and typed into a word-processing database. The output was then printed. These comments also helped identify causes of problems and gave suggested solutions to them. In many cases, both the quantitative and qualitative data collected from the raters supported the data collected from examinees, clearly identifying problems with the tasks and in almost all cases suggesting remedies to those problems.

3. CONCLUSIONS

From our experience in developing the TOPT, we have seen how the use of these two instruments during field testing efficiently collects data that identify, from both the examinees' and raters' perspectives, poorly functioning tasks on a performance-based assessment. In addition, this method, by encouraging examinees and raters to provide focused written feedback, enables both to be involved in the test development process. These systematically obtained quantitative and qualitative data alert the test developers to specific problematic aspects of the test tasks or the test as a whole that need to be addressed. We found with the TOPT that in many cases the suggestions of examinees and raters could be used to improve the tasks on the test. Finally, this method of collecting field test data gives the test developers confidence in the quality of those tasks receiving high ratings by examinees and raters. Similarly, the method provides a rationale by which tasks which receive poor ratings and are beyond repair can be deleted from the test.

The information on task functioning collected through this process can be useful to test score users as well as test developers. Should the validity or fairness of the test be challenged in court or in other ways, the information provides proof that examinees' and raters' reactions to the tasks were sought out during the test development process. The information also provides proof that problematic tasks were either deleted or revised according to examinees' and raters' suggestions. The process also allows the test developer to demonstrate that tasks not revised or deleted were perceived to be fair, appropriate and useful measures by pretest examinees and raters. Ultimately, when a sample of

examinees indicates that a certain task "allowed me to give a response that reflects my current ability," evidence for the validity of that task is generated. Thus, the mean rating assigned by examinees to a task can serve as an index of ecological validity as perceived by the group most directly affected, the examinees.

Use of this method has additional benefits. For example, four parallel forms of the field testing version of the TOPT were developed simultaneously, each parallel in terms of the language functions addressed by its tasks. The use of the field testing method described in this article allowed for the comparison of the functioning of parallel tasks across forms. In some cases, when only one of a set of four parallel tasks received a lower rating, it was possible for us to identify and correct that aspect of the lower rated task that distinguished it from similar tasks. This increases the test developer's knowledge about task characteristics that may affect performance. Additionally, comparing mean ratings of parallel tasks across the four forms identified generic task types that in general were perceived most favorably, as well as those that were perceived to be more problematic to examinees than other task types. For example, for the 119 examinees who took the field test Spanish TOPT, the picture-based task requiring narration in the past tense received the highest mean overall rating across forms (3.90). The task requiring a summary of factual information received the lowest mean overall rating across forms (3.33). This information can be added to the test developer's storehouse of knowledge with the result that he or she will know with some degree of confidence how both examinees and raters will perceive tasks of each type should they be used in the future.

We believe this method is both efficient and critical when the performance-based assessment consists of a single task, as in an essay test or writing sample. A Likert-type questionnaire addressing salient points can be developed and administered to field test examinees. After concluding the test, examinees can then be given time to reflect on their test-taking experience and directly describe or document problems they had with understanding the task, organizing their response, or with the amount of time they were allotted. This is more accurate and efficient than having readers try to infer what the deficiencies of a task may have been on the basis of a reading of the examinee's essay, which is currently the common practice.

Indeed, not all of the pretest papers need to be read in order to obtain data on the adequacy of the test or the task with our method. Since raters' honoraria is often the largest cost in performance-based assessment, during field testing more data can be collected for less cost by having examinees describe and document problems with the tasks that comprise the test. Field test raters can then rate a random subset of the essay papers, commenting on the adequacy of the task to elicit a ratable writing sample, without losing the input into the field testing process of those examinees whose papers were not read by the raters.

We employed this procedure (rating a random subset of the tests) with noteworthy success. The TOPT field test raters listened to only a subset (67%) of the Spanish field test tapes and evaluated the quality of each task by indicating on a scale of 1 to 3 the task's usefulness in determining the appropriate rating. While in our experience raters' comments were useful, examinees' comments were even more helpful and precise, and could be collected without incurring the cost of rating the tape. Thus, we feel that while both examinees' and raters' comments should be sought during the field testing of performance-based tasks, the examinees' comments are more useful and can be obtained more efficiently. Currently few performance-based tests systematically seek out examinees' comments during the test development phase. Our experience with the TOPT suggests that paying attention to examinee comments can go a long way toward perfecting the final test product.

REFERENCES

ACTFL. 1986. *ACTFL proficiency guidelines*. Hastings-on-Hudson, NY: American Council on the Teaching of Foreign Languages.

Coffman, W.E. 1971. Essay examinations. In R.L. Thorndike (Ed.), *Educational measurement*. Washington, DC: American Council on Education.

Haynes, S.N. & Wilson, C.C. 1979. *Behavioral assessment*. San Francisco: Jossey-Bass.

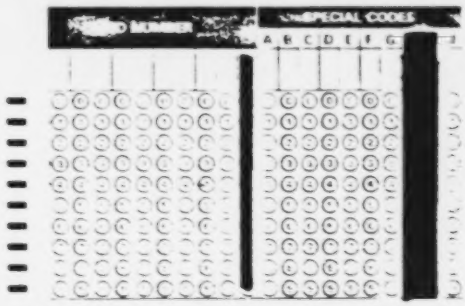
Priestley, M. 1982. *Performance assessment in education and training: Alternative techniques*. Englewood Cliffs, NJ: Educational Technology Publications.

Siegel, A.I. 1986. Performance tests. In R.A. Berk (Ed.), *Performance assessment: Methods and applications*. Baltimore, MD: Johns Hopkins University Press.

Stansfield, C.W. 1989. *Simulated oral proficiency interviews. ERIC Digest*. Washington, DC: ERIC Clearinghouse on Languages and Linguistics. (ERIC Document Reproduction Service No. ED 317 036)

Stiggins, R.J. 1987. NCME instructional module on design and development of performance assessments. *Educational Measurement: Issues and Practice*, 6, 33-42.

APPENDIX: Machine-readable Questionnaire for Obtaining Quantitative Data from Examinees for the TOPT Field Testing



GENERAL PURPOSE DATA SHEET II
form no. 70921

Name _____

Date _____

Please circle your answers

TOPT Language: FRENCH SPANISH

TOPT Form: A B C D

Your Subject: French Spanish bil ed

Strongly Disagree 1

Disagree 2

Agree and Disagree 3

Agree 4

Strongly Agree 5

USE NO. 2 PENCIL ONLY

1. The opening conversation helped ease me...---------->
2. Opening conversation time allowed was appropriate-->
3. Picture item #1 allowed a response that reflects--->
4. The map for picture item #1 was clear-------------->
5. Picture item #1 time allowed for preparing answer-->
6. Item #2 allowed response that reflects my ability-->
7. The drawing for picture item #2 was clear---------->
8. Picture item #2 time allowed for preparing answer-->
9. Picture item #3 allowed a response that reflects--->
10. The pictures for picture item #3 were clear-------->
11. Picture item #3 time allowed for preparing answer-->
12. Picture item #4 allowed a response that reflects--->
13. The pictures for picture item #4 were clear-------->
14. Picture item #4 time allowed for preparing answer-->
15. Picture item #5 a response that reflects ability--->
16. The pictures for picture item #5 were clear-------->
17. Picture item #5 time allowed for preparing answer-->
18. Topic item #1 allowed a response that reflects----->
19. Topic item #1 time allowed for preparing answer----->
20. Topic item #2 allowed a response that reflects----->
21. Topic item #2 time allowed for preparing answer----->
22. Topic item #3 allowed a response that reflects----->
23. Topic item #3 time allowed for preparing answer----->
24. Topic item #4 allowed a response that reflects----->
25. Topic item #4 time allowed for preparing answer----->
26. Topic item #5 allowed a response that reflects----->
27. Topic item #5 time allowed for preparing my answer->
28. Situation item #1 allowed a response that reflects->
29. Situation item #1 time allowed for preparing answer>
30. Situation item #2 allowed a response that reflects->
31. Situation item #2 time allowed for preparing answer>
32. Situation item #3 allowed a response that reflects->
33. Situation item #3 time allowed for preparing answer>
34. Situation item #4 allowed a response that reflects->
35. Situation item #4 allowed for preparing my answer-->
36. Situation item #5 allowed a response that reflects->
37. Situation item #5 time allowed for preparing my---->
38. I was not unduly nervous during the test---------->
39. I would prefer a beep signal as a signal to begin-->
40. A person listening will get an accurate picture---->

FUZZY SET THEORY APPLICATION IN CRITERION-REFERENCED TESTING

Felly Stoyanova
Sofia University

The article illustrates the application of fuzzy set theory for determining content and decision validity and setting cut-off points for criterion-referenced tests. An 18-item test of mathematics was administered to 135 third and fourth graders, and several methods for setting standards were compared. The article is included in the book because its methodological approach might be of interest also for experts in language testing.

1. BASIC CONCEPTS IN FUZZY SET THEORY

Fuzzy set theory was created for modelling processes and relations in complex social systems where many subjective factors and ambiguous notions exist.

Successful attempts for its application have been made in many fields, such as architecture, linguistics, biology, medicine and so on, but to the best of my knowledge it has not been applied yet in the sphere of psychological and educational measurement and evaluation.

Obviously there are many possibilities for such application as many concepts in this area have not been precisely defined. Every specialist has his/her own understanding of the different topics and the degree of agreement with a certain decision is not always definitely "yes" or "no" but can be intermediate-valued. Fuzzy set theory envisages this and takes it into consideration.

According to ordinary set theory, a set is defined if we know its elements. Thus membership value is considered to be 1 for objects belonging to the set and 0 for objects which are not set members.

In fuzzy set theory, however, membership value may not be just 0 or 1; it may be any digit in the interval [0:1], according to our degree of certainty that a given object belongs to this set. Thus a fuzzy set (subset) A is said to be defined when the membership value to A is determined in some way for every element of a certain ordinary basic set X. (The term "fuzzy set" was used for the first time by L. Zadeh [in 1965].)

There are many different methods for determining the membership function. Thus for examples:

(a) Membership value is defined as a proportion of experts who accept that an object belongs to the ordinary set A;
(b) Membership value is determined on the basis of comparing pairs of objects;
(c) Membership function is determined on the basis of constructing several subsets out of the same fuzzy set A;
(d) Membership value is determined through the usage of standard functions.

We are not going to dwell on the above methods in detail as they have been considered in many books on fuzzy set theory.

For our purpose, however, we will introduce some basic notions:

A convex combination of certain fuzzy sets $A_1, A_2, ..., A_n$ on the ordinary set X is a fuzzy set B on X whose membership function $m_B(x)$ is determined after the following equation:

$$m_B(x) = \sum_{i=1}^{n} a_i \cdot m_{A_i}(x), \text{ where } \sum_{i=1}^{n} a_i = 1 \qquad (1)$$

A relative distance between two fuzzy sets, A and B, is a positive number d(A,B) which is determined by means of the following equation:

$$d(A,B) = \frac{\sum_{i=1}^{n} |m_A(x_i) - m_B(x_i)|}{n} \qquad (2)$$

The linear index of fuzziness of a given fuzzy set A, f(A), is determined according to the equation:

$$f(A) = \frac{2 \cdot d(A,A')}{n} \qquad (3)$$

where A' is an ordinary set, the one nearest to the fuzzy set A and whose membership function is determined as follows:

$$m_A'(x) = \begin{cases} 0, \text{ if } m_A(x) < .5 \\ 1, \text{ if } m_A(x) > .5 \end{cases} \quad (4)$$

A. Prade and D. Dubois used the scalar functions of evaluation for a fuzzy set and thus determined also some indices for comparison between fuzzy sets, such as the index of similarity, inclusion index, etc. (Прад. Ал и Д. Дюбуа, 1986, p. 12-16).

The last notion we would like to introduce is level g of a fuzzy set A - this is an ordinary set A_g whose elements have membership value to A not less than g.

If we consider the notions of criterion-referenced testing from the point of view of fuzzy set theory, we shall see that there are possibilities for many different interpretations and, what is more important, for creating new approaches to solve old problems.

Let us assume that a basic ordinary set X is a set of test items included in a criterion-referenced test. Then we could define the fuzzy subset A on the ordinary set X as a set of items corresponding to the performance domain about which the test results will be used to make inferences. If we then determine the membership value of each item to this fuzzy set, we would consider this value as index of its item validity.

An index of content validity of the whole test could be some of the scalar estimators, or fuzziness indices.

If more experts take part in the validation procedure, their expert assessment could be regarded as different fuzzy sets A_i on the same ordinary set X and the generalization of their assessments could be the convex combination of these fuzzy sets. In this way the A_i coefficients could be chosen with respect to the experts' competence.

If the test consists of equivalent items constructed on the basis of an item form, IOX-type specification, map sentences or another method producing equivalent items, the item validity of all items will be one and the same and the content validity of the whole test will be the same then.

If we consider the people examined with the test as elements of the ordinary set X and define the fuzzy set of people who are masters in a corresponding performance domain, and determining the membership value of every element of X to A, we could use this approach for setting a standard cut-off for this set and for determining the decision validity of the same test.

The purpose of this paper is to illustrate the application of fuzzy set theory for determining the content and decision validity and setting standard of criterion-referenced tests, and to analyse the effectiveness of these methods.

2. COMPARATIVE ANALYSIS

This part will deal with the results of some comparative studies on the basis of a test in mathematics for primary school students.

2.1. Content Validity

In test planning, a limited-focus strategy was used (Popham, 1980, p. 21-22) and the test was aimed at assessing students' achievements, for one objective only, in the fields of primary school mathematics.

Six different IOX-type test-specifications were prepared for this objective. The main difference between them consisted in choosing the test items formats.

Fourteen experts (primary school teachers and mathematicians) took part in a preliminary determination of the content validity of these six tests.

The experts had to answer the question: "Does a given test specification correspond to the objective?", regarding all six test specifications. The answer, however, had to be provided in different ways:

(a) By comparison in pairs for all combinations of six test specifications (15 pairs);
(b) By a binary ("yes" or "no") response;
(c) By using a rating scale (five points);
(d) By constructing different level g_i of the fuzzy set.

Six different methods of determining the content validity of the different test specifications were used, on the basis of the experts' answers. In this way six different indices of content validity for the test specifications were obtained (Table 1):

Table 1. Content Validity Indices for Six Test Specifications

i	A_i	L_i	P_i	T_i	F_i
1	.82	.57	.79	.57	.74
2	.88	.71	.86	.68	.82
3	.80	.57	.79	.52	.73
4	.71	.71	.86	.43	.70
5	.71	.29	.64	.38	.64
6	.77	.43	.71	.39	.72

A_i - indices received on the basis of results using rating scaling (Aiken, 1985)

All indices are statistically significant. The homogeneity of the estimations is also statistically significant for all test specifications.

L_i - indices of content validity determined according to the answers to the binary question. The formula for L_i was proposed by Lawshe (Bishkov, 1991);

P_i - membership values of the test specification to the fuzzy set X (the set whose elements belong to the corresponding behavioral domain).

These indices are also obtained on the basis of the experts' answers to the binary question. Their values are equal to the proportion of experts who thought that a test item corresponded to this behavioral domain. There is a very close connection between the last two indices L_i and P_i. This connection can be represented by the following equation:

$$P_i = \frac{L_i + 1}{2} \qquad (5)$$

T_i - The membership values T_i are determined on the basis of comparison in pairs. The procedure for determining them is similar to the one proposed by Thurstone (Borisov et all, 1980, p. 50).

F_i - These membership values to a fuzzy set X are obtained on the basis of Yager's method (Yager, 1986). For this purpose each expert had to construct 20 g_i levels to the fuzzy set X, where g_i = 0.5, .10, .15, ..., 1.

As we can see from the results in Table 1, there is a very distinct correspondence between different indices, regardless of the fact that some of them are ordinary indices of content validity (A and L), and others are membership values to the X fuzzy set (P, T and F).

Usually, A_i indices are the highest, and T_i the lowest. P_i and L_i indices are normally in the middle and, following Equation 5, $P_i > L_i$ for every i.

If we regard A, P, T and F as fuzzy sets on the basic ordinary set with six elements (the six test specifications), we can determine the relative distance between them and, on this basis, make conclusions about the connections between these four different approaches for determining content validity. (The L index is excluded as it varies from -1 to +1 and could not be interpreted as membership value.)

Conclusions regarding the connections between the different indices could also be made on the basis of the analysis of Spearman rank Correlations.
In Table 2, the main diagonal of the matrix are given Spearman coefficients, and under the main diagonal the relative distances.

BEST COPY AVAILABLE

Table 2. Matrix of Spearman Rank-difference Correlation Coefficients and Relative Distances Between A, P, T, and F

	A	P	F	T
A	x	.39	.97*	.89*
P	.06	x	.51	.71
F	.06	.05	x	.94*
T	.29	.28	.23	x

As we can see from this table, the coefficients between the fuzzy set F and other fuzzy sets are highest and also statistically significant. This set is the one nearest to other sets as well.

The above results show that:

- The different indices of content validity resemble each other closely;
- Index F is more preferable, as it is nearest to all other indices and, on the other hand, its coefficients of correlation with other indices are the highest and statistically significant.

The F index displays some problems:

- The procedures for experts' evaluation and determining membership values are more time-consuming, the latter requiring a computer (without a computer it would be not impossible, but very difficult, to process);
- Some experts have difficulties in understanding the procedures and using such a manner of evaluation because they are not used to expressing their degree of certainty with numbers between 0 and 1.

There are some other advantages, however, additionally to the ones we mentioned above:

- This method does not demand direct determining of the degree of belonging which is usually very difficult and could produce many imprecise results;
- This method allows for some mistakes and imprecise expert's estimation;
- The generalization of the results of different experts allows us to take into consideration the level of competence of the different experts;
- The reliability of this method is very high. The mean of the distance between two fuzzy sets on the same ordinary set X, obtained for one expert in two investigations (with one week difference in time), is 0.09. It is a very small distance and we can conclude that this method has a good reliability over time.

2.2. Standard Setting and Decision Validity

This part will deal with the results of setting standards and determining decision validity for a test in mathematics which consists of 18 equivalent items. (The test was prepared on the basis of the third test specification which was considered in the above.)

This test was given to 135 examinees in the third and fourth grades.

For all examinees, the experts (their teachers) determined their membership value to the fuzzy set M, which was defined as a fuzzy set of the masters in this behavioral domain. Also the teachers divided examinees into three groups: masters, non-masters and borderline-masters group.

Several methods were used for setting standards:

(a) The Borderline group Method: In this way a cut-off score of 11.23 was received;
(b) The Method of Contrast Groups: In this case the cut-off score was 9.59;
(c) The Angoff Method - The same 14 experts took part in the procedure and the obtained cut-off score was 11.62;
(d) All experts were asked the question "What is the minimal number of items a student has to solve correctly to be classified as master?" The median of their answers was 8.

Regardless of the cut-off scores received, several indices of decision validity were determined for all possible cut-off scores (1, 2, 3, ..., 18) and they are presented in Table 3, as follows:

d_i - relative distance between the fuzzy set M and ordinary set S (for which the membership value for each student is 0 if the student score is less than cut-off i, and 1, if the student score is greater than or equal to i);
r - the point biserial coefficient between the student's membership value and test score (0 or 1), when the cut-off is i;
P_i - proportion of correctly classified students according to the test, when cut-off score is i;
F_i - correlation coefficient between masters-nonmasters and pass-fail the test.

In Table 3 the different indices are shown for all possible cut-off scores.

Table 3. Decision Validity Indices for Different possible cut-off scores

cut-off score	d_i	r	P_i	F_i
1	.58	.27	.52	.25
2	.58	.27	.52	.25
3	.52	.38	.59	.37
4	.52	.38	.59	.37
5	.42	.53	.70	.53
6	.40	.50	.70	.47
7	.40	.50	.70	.47
8	.35	.54	.74	.50
9	.32	.59	.78	.56
10	.32	.59	.78	.56
11	.30	.64	.81	.63
12	.27	.67	.81	.62
13	.29	.54	.78	.56
14	.26	.64	.81	.66
15	.27	.61	.70	.47
16	.27	.61	.70	.47
17	.27	.61	.70	.47
18	.29	.58	.63	.32

As we can see, there is a very close connection between different indices of decision validity for one and the same cut-off. This conclusion follows also from the analysis of the coefficients of product-moment correlation between the values of the different indices (Table 4). All correlation coefficients are statistically significant and show that the different indices are indicators of one and the same, or, to be more precise, indicators of decision validity. Which of them will be chosen is entirely dependent on us.

Table 4. Matrix of correlation coefficients between the indices d, rpb, P, F

	d	rpb	P	F
d	x	-.95	-.81	-.69
rpb		x	.87	.79
P			x	.97
F				x

Regardless of our choice, however, it would be better in setting standards to have indices of decision validity for all possible cut-off scores and, on this basis, make conclusions about the test standard. Thus, for example, as previously said, different methods of setting standards were used and, correspondingly, different cut-offs were received. If we look at Table 3, we will see that best cut-off score with respect to decision validity would be cut-off score = 12. That means that the most effective in this case are the Borderline-Group Method and Angoff Method. A good decision from this point of view would be also to accept 14

points as a cut-off score, regardless of the fact that this value was not received by any of the four methods used to set standard.

Another problem in determining decision validity is when we can accept that our index of decision validity is high enough. If we use the indices d or P as decision validity indices, we would not be able to give a precise answer. The statistical significance of F and r could be used as such criteria. But if we look at Table 4, we will see that the F index correlates with other indices relatively less than r does.

For this reason, perhaps the best solution would be to use as a decision validity index the point biserial coefficient (or rbis) between a student's membership value and his test results (pass or fail) with respect to the chosen cut-off score.

In conclusion, we could say that only a small part of all possible fuzzy set theory applications was demonstrated in this paper, investigations being in their initial stage. The biggest advantage of this approach is its accessibility, the possibility for easy interpretation and the fact that it takes into consideration the subjective factor. It would be interesting to see the theory applied in foreign language testing as well.

REFERENCES

Aiken, L. 1985. Three coefficients for analyzing the reliability and validity of rating. *Educational and Psychological Measurement* 45, 131-142.

Popham, W. 1980. *Domain Specification Strategies. In: Criterion-referenced Measurement: The state of the art*. Ed. by R. Berk, The John Hopkins University Press, Baltimore and London, p. 15-31.

Popham, W. 1981. *Modern Educational Measurement*. Englewood Cliffs, N.J.: Prentice-Hall.

Беллман, Р., Л. Заде. 1976. Принятие решений в расплывчатих условиях. сб. Вопросы анализа и процедуры принятия решений. ред. И. Шахнов. М. Мир. с. 172-215.

Бижков, Б. Теория и методика на дидактическите тестове. С. (под печат).

Борисов, А. и др. 1982. Модели принятия решений на основе лингвистической переменной. Рига. Знание.

Бърнев, П. и П. Станчев. 1987. Размити множества. С. Народна просвета.

Заде, Л. 1976. Понятие лингвистической переменной и его применение к принятию приближенных решений. М. Мир.

Прад, А. и Д. Дюбуа. 1986. Общий подход к определения индексов сравнения в теории нечетких множеств. В: Нечеткие множества и теория возможностей. Последние достижения. Под. ред. Р. Ягера. М. Радио и связь. с. 9-21.

Ягер, Р. 1986. Множества уровня для оценки принадлежности нечетких подмножеств. В: Нечеткие множества и теория возможностей. Последние достижения. Под. ред. Р. Ягера. М. Радио и связь. с. 71-78.

Section III

Empirical Investigations and Test Development

PERSONALITY CHARACTERISTICS AS A POTENTIAL SOURCE OF LANGUAGE TEST BIAS

Vivien Berry
Nagoya University of Commerce & Business Administration

The paper is concerned with the role personality plays in test performance. Two studies on Japanese college students suggest that extreme extraverts and introverts differ in how well they perform on certain language tasks, such as filling in a text on the basis of visual and aural stimuli, and performing in oral tests. The type of oral test (individual interview, homogeneous vs. heterogeneous pairing in paired interview) appeared to produce a slightly different effect depending on the personality type.

Recent research into individual differences in learners' cultural background (Chen & Henning, 1985; Zeidner, 1986, 1987), prior knowledge (Alderson & Urquhart, 1985), gender and status (Porter, 1990), and related issues, has shown that such factors can affect performance on language tests. In addition, there is some evidence that differences in cognitive characteristics such as tolerance of ambiguity (Chapelle & Roberts, 1986), and degrees of field independence (Hansen & Stansfield, 1981; Stansfield & Hansen, 1983; Hansen, 1984; Chapelle, 1988; Chapelle & Green, 1992) may also effect performance on certain types of language tests. However, as Bachman (1990) points out:

> "There is no evidence as yet relating performance on language tests to other characteristics such as inhibition, extroversion, aggression, attitude, and motivational orientation, which have been mentioned with regard to second language learning." (Bachman, 1990:275)

Unfortunately, major reviews of the role played by personality variables in second-language learning (Ellis, 1985; Skehan, 1989) have reached extremely pessimistic conclusions, particularly with regard to the implications of extraversion as a variable. They point out that many studies have failed to produce any significant findings, citing, for example, Naiman et al (1978), who did not find a significant effect for extraversion in characterizing the good language learner. Another study, described by Brown as "...the most comprehensive study to date on extraversion." (Brown 1987:110) is that of Busch (1982), who also failed to find support for her hypothesis that "extraverts are

more proficient in English." (Busch, 1982:109). Such results have led a number of L2 researchers to question the validity of the personality scales developed for psychological use. Ellis, for example, suggests that "...a failure to find an expected relationship (e.g. between extraversion and proficiency) may be because the test used to measure the personality trait lacks validity." (Ellis, 1985:122). Skehan is even more skeptical in his statement: "...there are grounds for questioning the desirability of adopting, wholesale, a construct from a feeder discipline, psychology, rather than subjecting the construct to further analysis to relate it specifically to language learning." (Skehan, 1989:105).

It can, however, be argued that the problem lies not so much with the lack of significance of the results obtained (also see Guttman, 1971; Cantor, 1985, for an interesting reappraisal of the value of "significance" in psychological research), but from the fact that previous L2 research has not, in reality, adopted wholesale constructs from psychology, but rather adapted them to test things "which intuitively strike them as important." (Ellis, 1985:120). It may also be that these pessimistic conclusions have been reached through reviewing research which tested hypotheses that are neither logically derived from personality theory, nor predicted from relevant experimental evidence. Summarizing a comprehensive review of second language personality studies, Griffiths (1990) concludes:

> "...the fact that researchers have not found relationships cannot be fairly used (as it has been) to dismiss personality variables from the research agenda; nor can highly validated psychometric instruments be held accountable for the failure." (Griffiths, 1990:5)

This paper will present evidence from two studies which suggests that when an appropriate instrument is used to assess personality, and when theoretically sound hypotheses derived from the psychological literature are tested, significant differences can be observed in the responses of introverts and extraverts on particular tasks.

A number of instruments have been developed which attempt to measure the major dimensions of personality, amongst them Cattell's 16PF (Cattell et al, 1970) and the Minnesota Multiphasic Personality Inventory (MMPI). Neither of these has been validated for use in Japan and one of the major difficulties is that concepts like Introversion-Extraversion (common to all of them) which have "... an agreed meaning in one culture may not have the same, or indeed any, meaning in another culture." (Iwawaki et al, 1980). The psychometric instrument used to assess degrees of extraversion in both studies reported here was the 86-item Japanese version of the EPQ (Eysenck Personality Questionnaire, Eysenck & Eysenck, 1975) which emerged from the cross-cultural validation studies carried out by Iwawaki et al in 1980. The importance of cross-cultural validation studies is pointed out by SBG Eysenck who argues that "... it is imperative that all items be tested for appropriateness before inclusion in any foreign scoring key", whilst warning of the dangers of "spurious results" if this

is not done (Eysenck, 1983:381). Methods of validation of the EPQ have been criticised, most notably in the methods used to derive indices of factor comparison (Bijnen et al, 1986). A full discussion of the criticisms, defenses and development of validation procedures of the EPQ is beyond the scope of this paper. However recent studies (Hanin et al, 1990) accept these criticisms and now state their results not in terms derived solely from factor analysis, but also from multi-dimensional scaling using smallest space analysis (Lingoes, 1973).

The major personality dimensions are represented in almost all large scale studies and nearly all theoretical formulations. They are represented by continua, the extremes of which can be described through idealized types:

> "Extraverts are sociable, like parties, have many friends and need excitement; they are sensation seekers and risk-takers, like practical jokes and are lively and active. Conversely introverts are quiet, prefer reading to meeting people, have few but close friends and usually avoid excitement. (Eysenck & Chan, 1982:154)

A review of the experimental research reported in the psychological literature reveals several studies which could prove fruitful areas for investigation in a second-language learning context. Two small scale studies which attempt to relate findings from psychological research to second language testing are reported here.

1. STUDY 1

The first study attempted to investigate the finding of Eysenck that "extraverts opt for speed, introverts for accuracy, when there is the possibility of a choice in the carrying out of an experimental task." (Eysenck, 73:205). Numerous replications of his study produced essentially the same results. More recently, Howard & McKillen (1990) also found that extraverts were fast, inaccurate and unreflective on a maze test, whereas introverts were slow and accurate. It was therefore hypothesized that introverts would show improved performance over extraverts on an English language test which required careful and accurate listening for detail.

The 86-item Japanese EPQ (Iwawaki et al, 1980) was administered to 130 female, second-year junior college students. Means and standard deviations obtained on the E scale for the college sample were very similar to those reported by Iwawaki et al.

In addition, a battery of language tests was also administered to identify differences in language proficiency. Results of the EPQ and of the language tests subsequently identified 43 extreme introverts or extraverts with no significant differences in language proficiency (as measured by a composite of results

obtained from a) an Institutional TOEFL test, b) an in-house placement test consisting of two cloze passages and c) the university English entrance exam.) 21 students were identified as being extreme extraverts (mean 17; s.d. 1.04; range 16-20) and 22 as extreme introverts (mean 7; s.d. 1.8; range 1-8). `Extreme' was taken to be plus or minus 1 standard deviation from the mean.

Table 1. Comparison of means & standard deviations.

	mean	s.d.
Iwawaki et al. (1980)	11.4	4.47
Junior College sample	11.97	3.97

These 43 students were then given a test which required them to reconstruct a gapped text as they watched and listened to two short video sequences (20 gaps). The reconstruction required was at the short phrasal level, a maximum of four words being required in any given gap. Each video sequence was shown twice and time was given between and after the video presentations to allow for writing. Post-test enquiries revealed that students did not feel time pressure had effected their performance.

Reliability of the test was adequate for the purpose (alpha = 0.86). Analysis of results showed that although no significant differences in overall language ability between extreme Es and Is had been observed on the language test battery, results on the test described above, show that introverts performed considerably better than extraverts. T-tests revealed these differences to be significant (p=.03).

Table 2. Comparison of scores on listening/gapfill test.

	Introvert	Extravert	t	sig.
mean	17.90	14.85	2.234	0.031
s.d.	3.44	5.17		
n	.22	21		

Subsequent inspection of students' scripts showed that extraverts had a tendency merely to write down a key word from the phrase, either a vocabulary item or a verb stem, whereas introverts tried to be very accurate, often making several attempts to correct spelling. (Since the test was intended to test accuracy in listening and not in writing, spelling was not one of the criteria for scoring unless mis-spelling changed the meaning of the response or could simply not be understood). The hypothesis that extraverts would be more impulsive and less accurate was thus confirmed.

2. STUDY 2

The second study attempted to examine differences in oral ability as measured by scores obtained from an oral interview. The current practice in many oral interview situations, particularly the current Cambridge/RSA communicative tests, is to conduct one or more sections of the oral interview component of the tests in pairs. This study investigated the hypothesis, derived from Leith (1974), that there would be significant differences in performance of both introverts and extraverts on an oral interview, *dependent on method of pairing*. Specifically, it was predicted (again from Leith, 1974) that introverts would perform best if interviewed in homogeneous pairs, next best if interviewed as individuals and worst if interviewed in heterogeneous pairs. Extraverts, on the other hand, would again perform best in homogeneous pairs but would do next best in heterogeneous pairs and worst as individuals.

Table 3. (from Leith,1974) Achievements of students learning in homogeneous of heterogeneous personality pairs or as individuals.

	Methods:		
	Homogeneous pairs	Heterogeneous pairs	Individuals
Personality:			
Introverts	32.2	27.3	30.0
Extraverts	30.6	27.7	25.4
Significance of differences:	n.s.	n.s.	$p<.01$

Homogeneous vs heterogeneous pairs: $p<.01$
Homogeneous pairs vs individuals: $p<.025$

From the 43 students identified for the first study, 36 students (18 each of extreme E and I) were selected and consented to take part in the second study. They were randomly assigned to each of the three possible personality pairings and 24 interviews were conducted as follows: 6 individual interviews of both I and E (12), 3 homogeneously paired interviews of both I and E (6) and 6 heterogeneously paired interviews (6). This allowed for a total of six sets of scores to be analysed in each of the six possible categories (I-individual, E-individual, I+I, E+E, I+E, E+I). Interviews of the different categories were also conducted in random order.

Table 4. Methods of pairing for interviews.

Type	Number of interviews	Number of students
Individual I	6	6
Individual E	6	6
I + I	3	6
E + E	3	6
I + E	6	12 (6I + 6E)
TOTAL	24	36

The test itself consisted of a four part, highly structured interview designed to approximate the level and format of the Cambridge PET test. All pairs were given exactly the same test but the stimuli were varied slightly for the individual student interviews. The first part of all interviews was a warm-up phase designed to elicit personal information; part 2 consisted of a picture stimulus to elicit descriptions and recognition and discussion of differences; part 3 was an extension of part 2, intended to elicit comments on personal likes and dislikes. The final part consisted of a role-play, again prompted by a picture. All components were specifically designed to be relevant to the students' own experience and knowledge. This was, of course, in order to test their language ability and not their knowledge of British (or American) culture.

Table 5. Structure of interviews.

Part 1: Warm-up to elicit personal information
Part 2: Picture stimulus / description recognition and discussion of differences
Part 3: Extension of above / likes & dislikes
Part 4: Role-play (picture prompt)

The interviews were conducted by two experienced oral examiners. Attempts were made to control for as many other potential sources of bias as possible in terms of examiner-examinee interaction. Both interviewers were female (as in the previous study, all students were female junior college students), roughly the same age, had the same native background (British), and had the same perceived status. The testing environments were almost identical and neither of the interviewers knew, or had taught, any of the students they interviewed. In addition, neither of the interviewers was aware of the personality characteristics of any of the individuals or pairs. Bearing in mind the nature of the study, both interviewers also completed the British version of the EPQ and were both found to be within the average range for their age and profession on the major scales.

Each interview was videotaped using unobtrusive, but not hidden, equipment. Despite the care with which the equipment was set up, and the fact that

identical equipment was used by both interviewers, subsequent differences in audio quality could not be controlled for. Retrospective viewing also revealed that despite having had many standardization of interviewing techniques meetings, differences were observed between interviewers in terms of accommodation of interviewees (Ross & Berwick, 1990) and in speech rate and response wait-time. Obviously these variables will need to be controlled for more rigidly in future replications. However, it is hard to see how the "human dimension" can altogether be removed from the situation.

All interviews were independently scored by six raters, three male and three female. One of each sex were categorized as extreme E, extreme I and within the average range respectively, on the British version of the EPQ. (All raters were British.) Criteria for scoring the interviews were described on a six point scale ranging from 0-5 for each of grammaticality, vocabulary, pronunciation and successful task completion. There were no significant differences in the ratings of males and females and, perhaps surprisingly, no significant differences between the ratings given by the different personality types. Students were allocated an average score in each of the categories and an overall average (normally the only score actually reported). Means of overall averages were calculated for each of the categories of interviews. Analysis of means yielded the following results:

Table 6. Comparison of scores on oral interview tests.

		Methods:	
	Individuals	Homogeneous pairs	Heterogeneous pairs
Personality:			
Introverts	61.15	69.80	68.33
Extraverts	56.04	80.21	71.35
Significance of differences:	n.s.	$p<.05$	n.s.

Introverts: Homogeneous vs heterogeneous pairs: ns
Individuals vs both pairs: $p < .05$
Extraverts: All results $p<.05$

As can be seen from the above results, the original hypotheses received some support. Extraverts performed exactly as predicted, showing dramatic increases over individual performance when interviewed in pairs and performing best of all in homogeneous pairs. Scores for introverts are highest in homogeneous pairs although these are not significantly different from those of heterogeneous pairs. However, against expectations, scores on individual interviews are significantly lower than on either of the pairings, suggesting that variables other

than extraversion are having an effect. In fact, as both introverts and extraverts do least well in an individual interview, it may be that culturally stereotypic views of the interviewer-interviewee relationship is disturbed by, for example, having to interact in a role-play situation. Nevertheless, given the size of the sample, it is quite surprising that even one personality type shows such clear cut results.

One fact which emerges from this study would be hard to dispute. Given the differences in scores obtained with different patterns of interactions, it would be extremely difficult to say with any degree of confidence which score (if any) represents a learner's true oral ability level. Obviously, the preliminary findings of such small-scale, localized studies can not be considered to provide conclusive evidence of bias in favour of a particular personality type on either of the tests they examined. Until they can be replicated on a much larger scale, possibly with learners from different cultural backgrounds, they can only suggest potential problems of interpretation of scores. However, given the pessimistic conclusions drawn from earlier reports of research into personality characteristics, the findings of these two studies surely at least suggest that such research is worthy of continued and sustained investigation.

****

This paper is dedicated to the memory of Roger Griffiths (1946-1991) whose encouragement and assistance with both studies was invaluable.

****

REFERENCES

Alderson, J.C. & Urquart, A. 1985. The effect of students' academic discipline on their performance on ESP reading tests. *Language Testing* 2,2.(192-204)

Bachman, L.F. 1990. *Fundamental Considerations in Language Testing*. Oxford. Oxford University Press.

Bijnen, E.J., Van der Net, T.Z. & Poortinga, Y.H. 1986. On cross cultural comparative studies with the Eysenck Personality Questionnaire. *Journal of Cross-Cultural Psychology* 17. (3-16)

Brown, H.D. 1987. *Principles of Learning and Teaching*. Englewood Cliffs, NJ. Prentice-Hall.

Busch, D. 1982. Introversion-extraversion and the EFL proficiency of Japanese students. *Language Learning* 32. (109-132)

Cattell, R.B., Eber, H.W. & Tatsouka, M.M. 1970. *Handbook for the Sixteen Personality Factor Questionnaire (16PF)*. Champaign, Ill. Institute for Personality and Ability Testing.

Chapelle, C.A. 1988. Field independence: a source of language test variance? *Language Testing* 5,1. (62-82)

Chapelle, C. & Roberts, C. 1986. Ambiguity tolerance and field dependence as predictors of proficiency in English as a second language. *Language Learning* 36. (27-45)

Chapelle, C. & Green, P. 1992. Field independence/dependence in second language aquisition research. *Language Learning* 42/1. (1-37)

Chen, Z. & Henning, G. 1985. Linguistic and cultural bias in language proficiency tests. *Language Testing* 2,2. (155-63)

Eilis, R. 1985. *Understanding Second Language Aquisition*. Oxford. Oxford University Press.

Eysenck, H.J. 1973. *The Inequality of Man*. London. Maurice Temple Smith.

Eysenck, H.J. & Eysenck, S.B.G. 1975. *Manual of the Eysenck Personality Questionnaire*. London. Hodder & Stoughton.

Eysenck, S.B.G. 1983. One approach to cross-cultural studies of personality. *Australian Journal of Psychology* 35 (3). (381- 91)

Eysenck, S.B.G. & Chan, J. 1982. A comparative study of personality in adults and children: Hong Kong vs England. *Personality and Individual Differences*, 3. (153-160)

Griffiths, R. 1990. *Language Learning and the Japanese Personality: Hypotheses from the Psychological Literature*. Paper presented at the IUJ Conference on Second Language Research in Japan. 20th October 1990. Tokyo.

Guttman, L. 1977. What is not what in statistics. *The Statistician*, 26. (81-107)

Hanin, Y., Eysenck, S.B.G., Eysenck, H.J. & Barrett, P. 1991. A cross-cultural study of personality: Russia and England. *Personality and Individual Differences*, Vol.12, No.3. (265-71)

Hansen, L. 1984. Field dependence-independence and language testing: evidence from six Pacific Island cultures. *TESOL Quarterly* 18. (311-24)

Hansen, J. & Stansfield, C. 1981. The relationship between field dependent-independent cognitive styles and foreign language achievement. *Language Learning* 31. (349-67)

Howard, R. & McKillen, M. 1990. Extraversion and performance in the Perceptual Maze Test. *Personality and Individual Differences*, Vol.11, No. 4. (291-96)

Iwawaki, S., Eysenck, S.B.G. & Eysenck, H.J. 1980. Japanese and English personality structure: a cross-cultural study. *Psychologia* 23. (195-205)

Leith, G.O. 1974. Individual differences in learning: Interactions of personality and teaching methods. Personality and Academic Progress proceedings. London. Association of Educational Psychologists. (14-25)

Lingoes, J.C. 1973. *The Guttman-Lingoes nonmetric program series.* Ann Arbor. Mathesis.

Naiman, N., Frohlich, M., Stern, H.H. & Todesco, A. 1978. *The Good Language Learner.* Research in Education Series #7. Toronto. Ontario Institute for Studies in Education.

Porter, D. 1990. Affective Factors in the Assessment of Oral Interaction: Gender and Status. In: Anivan, S. (Ed) *Current Developments in Language Testing.* Anthology Series 25. SEAMEO, Singapore. (92-102)

Ross, S. & Berwick, R. 1990. *The Discourse of Accommodation in Oral Proficiency Interviews.* Paper presented at the 25th RELC Regional Seminar, Language Testing and Language Programme Evaluation. April 1990.

Skehan, P. 1989. *Individual Differences in Second-Language Learning.* London. Edward Arnold.

Stansfield, C. & Hansen, J. 1983. Field dependence-independence as a variable in second language cloze test performance. *TESOL Quarterly* 17. (29-38)

Zeidner, M. 1986. Are English language aptitude tests biased towards culturally different minority groups? Some Israeli findings. *Language Testing* 3,1. (80-95)

Zeidner, M. 1987. A comparison of ethnic, sex and age biases in the predictive validity of English language aptitude tests: some Israeli data. *Language Testing* 4,1. (55-71)

RELATIONSHIPS BETWEEN TOEFL SUBSECTIONS, SPEAKING, AND LANGUAGE LEARNING BACKGROUND

Harry L. Gradman and Edith Hanania
Indiana University

The article reports the results of a study investigating the relationship between language learning background and English proficiency as measured by the TOEFL and the ACTFL Oral Proficiency Interview tests. The results indicate that reading outside class has a dominant and direct effect on all aspects of language proficiency, including listening and speaking.

At the Center for English Language Training at Indiana University we have been investigating the relationship between aspects of the language learning background of students in our Intensive English Program and their initial performance on the Test of English as a Foreign Language (TOEFL), which is commonly used for admission to university study in the United States. Our work so far has indicated that the background factor that has the strongest relationship to TOEFL is reading outside class, or extracurricular reading. Surprisingly, the extent of oral exposure to English and its use for communication in and out of the classroom were found to be only indirectly related to performance on the TOEFL, and indeed listening outside class, or extracurricular listening, showed a negative relationship with TOEFL scores (Gradman and Hanania 1991).

These results gave rise to several questions, foremost among which was whether background factors would relate in the same way to a holistic speaking test as they do with the written TOEFL. We have therefore extended the work by collecting background data from a new sample of 125 students to whom both the TOEFL and the Oral Proficiency Interview developed by the American Council on the Teaching of Foreign Languages (ACTFL) were administered soon after the students joined our program. The data were analyzed using correlation, multiple regression, and path analysis. The purpose was to find out the following:

1. How language learning background factors relate to each of three subsections of the TOEFL: listening comprehension, structure and written expression, and vocabulary and reading comprehension;

2. How these same factors relate to a test of speaking ability, such as the ACTFL Oral Proficiency Interview;
3. How scores on different aspects of language proficiency (listening, structure, reading, and speaking) relate to each other for the whole sample and for groups of students with different first-language backgrounds.

1. PROCEDURE

The subjects for this study were 125 students who were newly enrolled in the Intensive English Program at Indiana University. They came from a variety of first-language backgrounds, about equally distributed between Arabic, Japanese, Romance, and other languages. They had entered the program at different levels, depending on their language ability; and their initial TOEFL scores with us were normally distributed.

Information about the language learning background of the students was elicited by interview through an oral questionnaire. This background information concerned the students' learning of English before they came to us and fell into four major categories: formal learning of English, exposure to and use of English in class, exposure to and use of English outside class, and attitudes and motivation. The items of information were coded and quantified, resulting in over forty variables[1]. Based on intercorrelation and distribution analysis, as well as on the results of our previous work, five main variables were selected for further analysis in the present study: two classroom and three extracurricular variables. The classroom variables were: oral exposure to English in class or *listening in class*, and oral use of English in class or *speaking in class*. The extracurricular variables were: *listening outside class*, *reading outside class*, and *speaking outside class*. We considered these to be basic underlying variables that could provide us with information about important characteristics in the language learning background.

The measures of proficiency used in this study were the Test of English as a Foreign Language (TOEFL), and the Oral Proficiency Interview (OPI). The TOEFL, as is generally known, is a standardized instrument for language assessment which is widely used by universities in the United States for admission purposes. The multiple-choice examination consists of three sections: Listening Comprehension, Structure and Written Expression, and Vocabulary and Reading Comprehension. These TOEFL sections will be referred to in this paper as Listening, Structure, and Reading. The students in our Intensive

[1] Detailed information about the variables and the procedure appears in Gradman and Hanania, "Language Learning Background Factors and ESL Proficiency," *MLJ* 75:1 (1991), 39-51.

English Program take the institutional version of the TOEFL at the end of every 7-week session. The students' initial TOEFL scores were used in this study; that is, the scores that they obtained at the end of the first session with us.

The OPI is a measure of speaking proficiency developed by ACTFL. It was administered to the students about halfway through their first session with us. The test follows set procedures which aim to determine the ability of a learner to communicate in English in a variety of oral interactions. The test was administered by a certified ACTFL interviewer.

Three main types of data analysis were used:

1. Multiple regression analysis for five background variables, using in turn the total TOEFL, each of the TOEFL subsections, and the OPI as the dependent variable.
2. Path analysis using the above variables, to determine direct and indirect effects.
3. Intercorrelations for the scores on the total TOEFL, TOEFL subsections, and the OPI. This was done first for the whole sample and then by first-language subgroups.

2. RESULTS

The results are presented in three sections. The first two concern the relation between language background factors and different facets of language proficiency, as reflected by the TOEFL subsections: Listening, Structure, and Reading, and by the test of speaking, the OPI. The third concerns the intercorrelations between the different proficiency measures.

2.1. Multiple Regression

Table 1 gives the results of multiple regression analysis for the five independent variables: *listening in class, speaking in class, listening outside class, reading outside class,* and *speaking outside class;* and each of five measures of proficiency as dependent variables: TOEFL, the TOEFL sections of Listening, Structure, and Reading, and the Speaking Test. The Table gives the single or multiple correlation coefficients for the variables selected by stepwise procedure (criterion p=.05).

As can be seen, *reading outside class* is the first factor selected in the multiple regression equation for each of the proficiency measures. With respect to TOEFL, this result confirms our previous findings. What is of note, however, is that outside reading appears to be important for all the other measures as well, including the tests of listening comprehension and of speaking. The relationship

with the speaking test is particularly significant as this test consists entirely of an oral interview and does not involve any reading. The prominence of this factor points to an important role for reading in language learning.

Table 1 also shows that *listening outside class* is selected in the equation for the dependent variables of total TOEFL, Structure, and Reading. However, an inverse relationship exists between this factor and the proficiency measures, as indicated by negative betas in the regression equation. In other words, it appears that a higher proficiency score is associated with a lower score on this variable. Again, this confirms our previous findings with regard to TOEFL, and it extends the findings to the two subsections of the test, Structure and Reading, though not to Listening and not to Speaking. The negative effects observed for extracurricular listening has been a puzzling aspect of the findings, which we plan to investigate further. To some extent, it may be related to the fact that, in scoring this variable, no distinction was made between active and passive listening. The scores included time spent hearing English without attempting to process the language, as in watching an English movie but depending on the native-language subtitles for the meaning.

Table 1. Multiple Regression Analysis (N=125)

Independent Variables: *listening in class, speaking in class, listening outside class, reading outside class, speaking outside class*

Dependent Variable:		R	R^2	Sign of Beta
Total TOEFL	*Reading outside class*	.34	.11	pos.
	+Listening outside class	.40	.16	neg.
TOEFL Listening	*Reading outside class*	.28	.08	pos.
TOEFL Structure	*Reading outside class*	.39	.16	pos.
	+Listening outside class	.46	.21	neg.
TOEFL Reading	*Reading outside class*	.25	.06	pos.
	+Listening outside class	.33	.11	neg.
Speaking Test	*Reading outside class*	.33	.11	pos.
	+Speaking outside class	.39	.15	pos.

One other variable which appears in Table 1 is *speaking outside class*, which was selected in the equation for the Speaking Test. It is not surprising that speaking practice outside of class should show a significant relationship with performance on a speaking test. However, noticeably missing from the multiple regression results are the classroom variables of *listening in class* and *speaking in class*, which would be expected to have an effect on language performance as well. In order to investigate possible indirect effects from these variables, path analysis was carried out using the statistical program LISREL.

2.2. Path Analysis

The main object of this analysis was to investigate the indirect as well as the direct effects of the selected background variables on each of the proficiency tests. A simple basic model was developed for this purpose consisting of one of the proficiency measures, each taken in turn, and the same background variables used in the multiple regression analysis: *listening in class, speaking in class, listening outside class, reading outside class,* and *speaking outside class*. The basic model for TOEFL is given in Figure 1. The paths within the model are unidirectional, as indicated by the arrows. They begin with the classroom variables, based on the assumption that the classroom is generally the starting point for foreign language learning, and that it precedes individual exposure to and use of language outside class. The assumption is also made that the extent of exposure to and use of the target language in class is largely determined by the teacher. The paths in the model end with the proficiency test, which is the end point in our study, in that all the background data pertains to the period preceding the taking of the test.

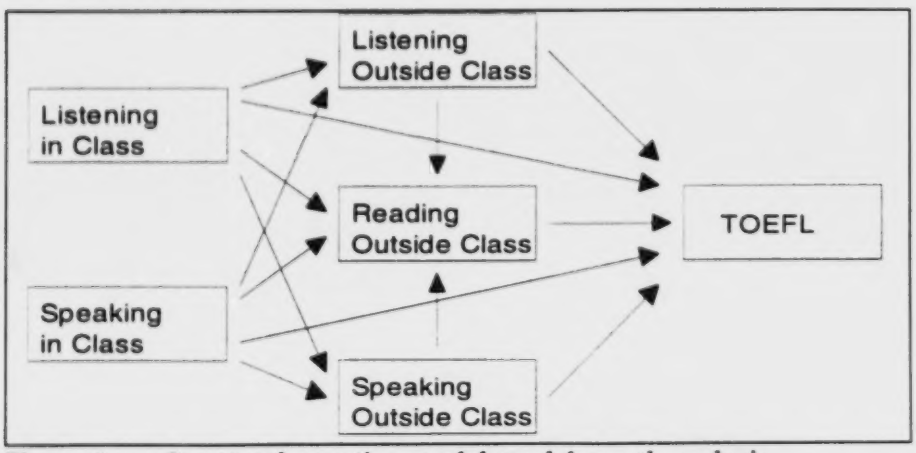


Figure 1. Structural equation model used for path analysis.

The analysis for each test starts with the full range of relationships shown in Figure 1. The model is then gradually adjusted and modified, eliminating paths where the effects are not significant (t-value of less than 2) and adding paths where necessary as indicated by a high modification index, until a model is found that best fits the data. This best fitting model in our estimation is presented here as the result of the analysis. It shows paths where the effects were significant and gives the coefficients for the relationships along these paths. The results of path analysis for each of the tests are presented in Figures 2-6. The measures of overall fit for each of the models are given in Table 2.

Table 2. Measures of Overall Fit for Each Model

Model	Figure	df	chi^2	p	GFI	AGFI	RMR
TOEFL	2	4	3.48	.481	.989	.958	27.602
Structure	3	4	4.48	.345	.986	.947	.929
Reading	4	4	3.53	.474	.989	.958	.872
Listening	5	5	5.39	.371	.983	.949	.674
Speaking	6	4	5.36	.252	.983	.937	.160

df=degrees of freedom; GFI=goodness-of-fit index; AGFI=adjusted goodness-of-fit index; RMR=root mean squared residual.

In looking at the figures, two general observations may be noted. The first is that in all the resulting models for this sample, the variable *listening in class* had to be eliminated, but a path had to be added from *listening outside class* to *speaking outside class*. The second observation is that the effects on the left side of the model are identical in all cases. The effects that do vary are those of the outside class variables on each of the proficiency measures.

Looking first at the constant part of the models, we find that there are moderate effects from *speaking in class* both to *speaking outside class* and to *listening outside class*, a smaller but still significant effect from *listening outside class* to *speaking outside class*, and a moderate effect from *speaking outside class* to *reading outside class*. What these paths indicate is that *speaking in class* and *listening outside class* have indirect positive effects on the proficiency measure through *speaking outside class* which in turn has a positive indirect effect through *reading outside class*.

Focusing on the variable components of the models, a comparison of the figures shows that the results are similar for TOEFL (Figure 2) and two of its subsections, Structure (Figure 3) and Reading (Figure 4). In each case, the coefficient showing the effect of *reading outside class* on the proficiency scores is moderately high, and there is a negative effect along the path from *listening outside class* to the test scores.

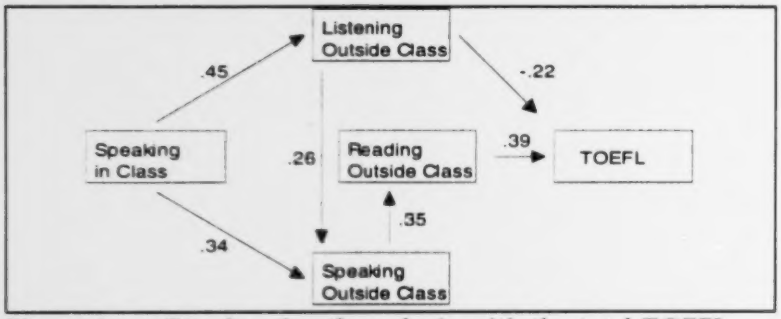


Figure 2. Results of path analysis with the total TOEFL.

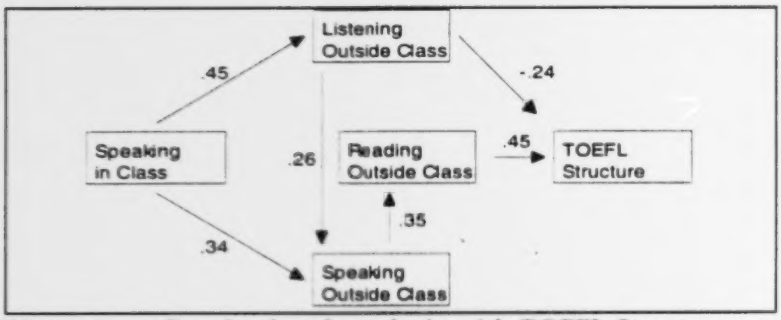


Figure 3. Result of path analysis with TOEFL Structure.

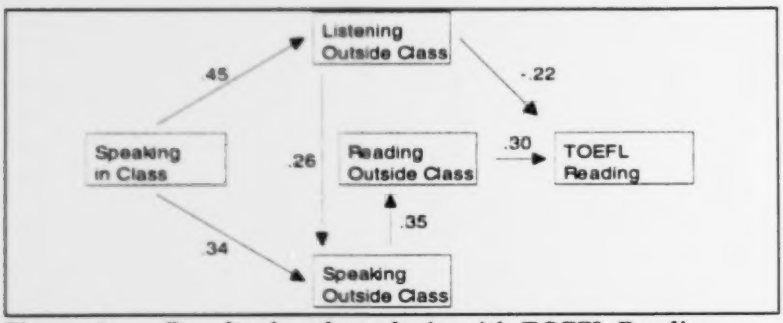


Figure 4. Result of path analysis with TOEFL Reading.

The results for TOEFL Listening (Figure 5) and for the Speaking Test (Figure 6) differ from the previous set. There is no path showing a negative effect between *listening outside class* and the test scores. In addition, there is a direct path with a positive effect from *speaking outside class* to the Speaking Test.

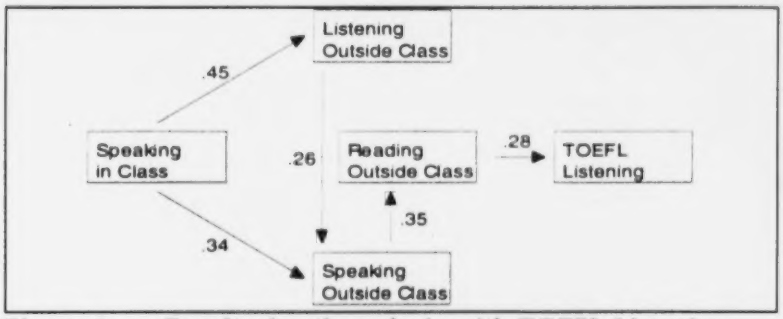


Figure 5. Result of path analysis with TOEFL Listening.

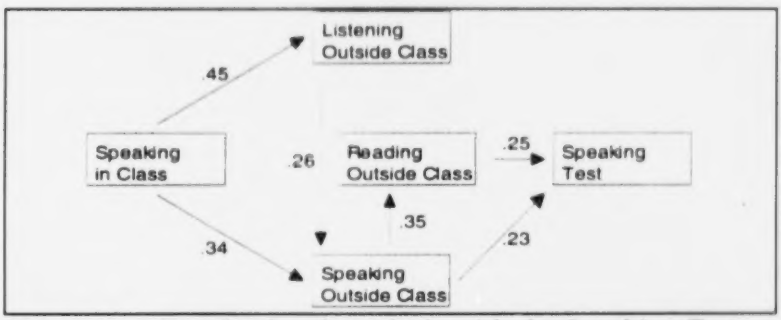


Figure 6. Result of path analysis with the Speaking Test.

The direct effects observed in the path analysis correspond with those seen in the multiple regression analysis. *Reading outside class* has a direct effect on the proficiency score in all cases. *Speaking outside class* has a direct effect on the speaking test, and *listening outside class* has a negative effect on proficiency scores, except for those of the oral skills of listening and speaking. What path analysis has revealed, however, is the indirect effects from exposure to and use of the target language in class and out of class through their effects on outside reading.

2.3. Intercorrelations

The intercorrelations between the TOEFL, its subsections, and the speaking test are presented in Table 3. Part A of the Table gives the intercorrelations for the whole sample. The intercorrelations for first-language groups are given for Arabic in Part B, Japanese in Part C, and Romance in Part D. A fourth group consisting of students from a variety of other languages was considered to be less meaningful and is therefore not included in Table 3.

Table 3. Intercorrelations

	TOEFL	Listening	Structure	Reading	Speaking
A. All (N=125)					
TOEFL	1.00				
Listening	.89**	1.00			
Structure	.94**	.78**	1.00		
Reading	.92**	.72**	.80**	1.00	
Speaking	.58**	.58**	.55**	.47**	1.00
B. Arabic (N=31)					
TOEFL	1.00				
Listening	.90**	1.00			
Structure	.90**	.72**	1.00		
Reading	.91**	.77**	.69**	1.00	
Speaking	.69**	.73**	.49**	.67**	1.00
C. Japanese (N=33)					
TOEFL	1.00				
Listening	.94**	1.00			
Structure	.96**	.87**	1.00		
Reading	.97**	.88**	.90**	1.00	
Speaking	.46**	.48**	.47**	.38*	1.00
D. Romance (N=30)					
TOEFL	1.00				
Listening	.91**	1.00			
Structure	.95**	.80**	1.00		
Reading	.89**	.69**	.80**	1.00	
Speaking	.62**	.62**	.66**	.40*	1.00

** Significant at .01
* Significant at .05

Focusing on the whole sample first (Table 3A), it is clear that the highest correlations are not unexpectedly between the total TOEFL score and its subsections Listening, Structure, and Reading. The next highest are between the TOEFL subsections of Structure and Reading. These intercorrelations are very similar to those published by the Educational Testing Service in the various editions of its *TOEFL Test and Score Manual*. Intercorrelations with the Speaking Test, however, are lower, though still significant. The lower correlations are to be expected since the skill being tested and the tasks involved are distinctly different in the oral interview from the TOEFL.

A comparison of part A of Table 3 with parts B, C, and D allows us to detect differences in the patterns of correlation for the three first-language subgroups. A few points may be noted. For the Arabic group, the correlations for speaking with the total TOEFL, listening, and reading are higher than they are for the whole group, and correlations of structure with reading and with speaking are somewhat lower. This suggests that, for the Arab group, their speaking scores reflect their overall proficiency as measured by the total TOEFL score better than they do for the whole group. On the other hand, their scores on structure do not reflect their overall ability as well as they do for the whole sample, or indeed for the other two language groups.

In the case of the Japanese group, the correlations of speaking with the total TOEFL and its subsections are distinctly lower than they are for the whole sample, and the correlations of structure with reading and listening are higher. In fact, for this group, all the intercorrelations between the total TOEFL and its subsections are relatively high. This implies that performance on any of these subsections would reflect overall proficiency as measured by TOEFL. However, the speaking ability does not seem to reflect performance on the TOEFL, and vice versa.

The third language group, Romance, includes first-language speakers of Spanish, French, Italian, and Portuguese. It is a less homogeneous group than either of the first two, and the correlation coefficients for this group are, in general, similar to those in the whole sample.

3. SUMMARY AND CONCLUSIONS

This study was carried out as an extension of a project investigating the relationship between aspects of the language learning background of students and their level of English proficiency, as reflected in performance on the TOEFL. Our previous work had indicated that extracurricular reading had a direct effect on TOEFL scores. In this study we explored the relation of selected language background variables with several aspects of language ability, as reflected in the

TOEFL subsections and in the speaking test OPI. We also sought to determine the intercorrelations between these separate tests.

Our results for language learning background factors indicate that reading outside class has a dominant and consistent direct effect on all facets of language proficiency, including listening comprehension and speaking. The results also show a negative relationship between listening outside class and the non-oral proficiency measures, a finding which needs to be pursued further. Results of the intercorrelation analysis for all the tests showed distinct differences in the patterns of relationship among the three language groups.

What are the implications of these findings for language teaching and testing? In relation to teaching, one implication is clear. Reading outside class appears to enhance language learning and should be encouraged in all language teaching programs. With respect to testing, two implications may be drawn. First, the results underline the need for using different kinds of assessment measures in evaluating language proficiency. Second, in interpreting the test scores, it is necessary to keep in mind that the interrelationships in students' performance on different tests may vary with first-language groups.

REFERENCES

Gradman, Harry L. and Edith Hanania. 1991. Language Learning Background Factors and ESL Proficiency. *Modern Language Journal*, 75(1), 39-51.

RECENT DEVELOPMENTS IN NATIONAL EXAMINATIONS IN FINLAND

Ari Huhta, Language Centre for Finnish Universities
Kari Sajavaara, Department of English
Sauli Takala, Institute for Educational Research
University of Jyväskylä

The article deals with two large examination systems currently in use in Finland, the Matriculation Examination, which is a school-leaving examination, and the Foreign Language Diploma for Professional Purposes, which is a language test for people working in business and administration. The structure and rating systems of the examinations are described.

INTRODUCTION

Generally speaking, the Finnish school system has until very recently been very centralized with national timetables, national curricula, centrally approved textbooks and a national system of teacher pre-service and in-service training. By contrast, assessment of pupil achievement and grading have been mostly internal in the Finnish school system.

There have never been any standardized tests in Finland. The only external examination in the school system is the long-established matriculation examination, which is a high stakes examination with a powerful washback effect on teaching in the upper secondary school (Konttinen 1993, Kuru 1993, Pasanen 1977; see also Sajavaara 1992). In addition to the matriculation examination, this paper will also deal with the Finnish Foreign Language Diploma for Professional Purposes, which has been designed as an examination to be recognized nationally.

1. THE MATRICULATION EXAMINATION

There is only one nationwide examination system in Finland, the Matriculation Examination. It goes back to more than a hundred years and was initially administered by the University of Helsinki as its own entrance examination. It now takes place at all senior secondary schools twice a year, in the spring and in the autumn.

The examination is administered, organized, and marked by the Matriculation Examination Board, which is appointed annually by the Ministry of Education. Its 30-odd members consist of representatives of Finnish universities and the National Board of Education.

The purpose of the examination is to find out if the students taking it have reached sufficient 'intellectual maturity' and master the content included in the official curricula. Thus the examination is intended to be a measure of both general ability and of what students have learned.

Compulsory tests include the mother tongue, the other national language, one foreign language, and either mathematics or general knowledge (including religion, psychology, history, physics and chemistry, biology and geography; students are now allowed to freely concentrate their answers on one or two of these subjects only). Optional tests (maximum of *two*) can be taken in mathematics, general knowledge, or two foreign languages.

The tests are constructed on the basis of the official curricula, and separate tests are administered for different curricula (representing programs of study which vary in terms of length of study and extent of content). For instance, three tests are available in the first foreign language and two tests in mathematics. Every time the tests are administered there are some protests from schools claiming that the tests have covered content not included in the curricula, but these criticisms are limited to only a few items.

There is a general rule that any changes in the examination set-up are announced three years in advance, so that upper secondary schools can make any adaptation to their teaching that they feel might be necessary.

1.1. The marking system

The students' answers are marked first by their own teachers (cf. defense m[illegible] in Shohamy's paper). They enter their mark on the test papers and on a separ[illegible]e list. All this material is sent to the Board, which uses the services of 300-350 raters to mark the papers. There are some 50 raters for marking the mother tongue essay and some 35 raters to do the same in English. Over the past few years, some 27,000 students have sat the examination each spring.

1.2. The mother tongue

The only kind of test used for Finnish/Swedish as a mother tongue is the written essay. Students are presented a set of 14 topics to choose from. The topics are related to what students have studied at school: religion and questions of value; culture; history and society; a person as an individual and as a member of society; nature, natural sciences and technological applications;

other topics. A few years ago students got an opportunity to write an essay on two occasions, and the teacher sends the better one for external assessment by the Board.

The test in the mother tongue has recently been partially revised after a long period of preparation and field testing. One of the two testing sessions is based on a variety of stimuli, mainly extracts from prose, fiction, and factual prose and presented in the form of a booklet. Students are expected to choose one of the given topics and write an essay drawing on the provided material in an appropriate manner. Topics are roughly similar to the ones described above.

The essays are rated on a scale of 0 - 100 and grouped into six categories from excellent ('laudatur') to fail ('improbatur'). The rating is based on a holistic marking scheme.

1.3. Second language/foreign languages (FL)

From 1874 until 1919 the FL tests consisted of translation from L1 to L2 with the help of a dictionary. From 1919 until 1965 the test was a translation test from L2 to L1 and from L1 to L2 with no dictionary. The text to be translated was limited to one A4 page. From 1965 until 1977 it was possible to choose an option of a more varied test, which from 1972 came to include multiple choice tests of reading and listening comprehension, and a composition. Since 1977 the three-part test was the only option available. In 1979 a 20-item fill-in type of cloze test of grammar was added.

In 1990 the Board started to implement a new phase, which makes it possible to vary the test types within a set of an approved list. There are some differences between languages, but in most language tests the first changes so far introduced have been the introduction of five open-ended questions in LC (students give brief answers in writing), and the earlier short fill-in cloze test of grammar has been replaced by a MC cloze integrated test of grammar, vocabulary, and idioms (35-50 items) which is based on unified texts.

There is a plan prepared by the National Board of Education, which is responsible for curriculum development, to introduce a component of oral testing. Pilot work has been going on for a couple of years in four 'regular' schools and in all university practice schools attached to teacher education departments. A collection of articles on oral proficiency and the testing of oral proficiency was published recently (Takala 1993).

1.4. Marking and grading

English/A & B language
Listening comprehension: 90 points; 20 MC items (weighted by 3), total 60 points; 5 open-ended questions (3 points weighted by 2), total 30 points
Reading comprehension: 110 points; 20 MC items (x 3), total 60 points; 50 MC cloze items, total 50 points.

Writing/essay: 99 points

The rating of essays is based on a holistic rating scheme, which is under revision. Open-ended answers to comprehension questions are rated on a scale 0 - 3. Criteria for rating summaries are being constructed.

1.4.1 Future Developments

A national working party submitted a report for the reorganization of the matriculation examination on June 21, 1993. The report suggests that four tests are obligatory, but they can be freely chosen by the students from among the following five subjects: mother tongue, the other national language, a foreign language, mathematics and a test of general knowledge. The student can supplement the examination with optional tests in foreign languages.

Students would be free to choose from two tests, a basic test and an advanced test, in mathematics, in the other national language and in the foreign language, but they have to take an advanced test at least in one of the three subjects.

The working party strongly emphasizes the importance of oral skills and recommends that oral testing be introduced as part of the matriculation examination. Initially, the oral test would be optional for students and it would be rated by teachers.

2. THE FINNISH FOREIGN LANGUAGE DIPLOMA FOR PROFESSIONAL PURPOSES

The Finnish Foreign Language Diploma for Professional Purposes (Työelämän kielidiplomi, TKD) was established to serve the sectors of working life in Finland which require an ability to communicate in foreign languages (see e.g. Sajavaara 1992). There was a need, most notably in Finnish business and industry, to create a nationwide system of certification for professional language competence. The University of Jyväskylä had carried out large scale foreign language needs assessments in the 1980s in certain key areas of Finnish society, and, partly because of that, was given the task of producing the language tests which make up the examination. The tests were designed by a group of experts

on LSP teaching and testing. The writers of the present paper are members of the group.

The diploma awarded on the basis of the examination serves two kinds of purposes: 1) individuals can obtain a nationally accepted certificate on their language proficiency in a professional context, and 2) employers can obtain an independent evaluation of candidates' language proficiency for selection purposes. The diploma contains both an overall grade and separate grades on speaking, listening, reading, writing skills plus grades on pronunciation, fluency, grammar, vocabulary, and the appropriateness of the candidate's language use. Each feature is assessed on a six-point scale; the diploma reports both numerical grades and short verbal descriptions of what they mean in terms of ability to use the language.

The examinations are advertised and administered by the Institute of Marketing and the Finnish Institute for International Trade (FINTRA) in Helsinki; the tests are developed and marked by the University of Jyväskylä. The first English examination was administered in 1989, the first German examination in 1991, and the first Swedish in 1992. The first Diploma tests have been geared towards the needs of people working in business and commerce, and in administration. Individual testees' language needs are analysed by questionnaires sent a few weeks prior to the examination; if the testee's needs appear to be markedly different from the typical needs in that field, some test tasks may be modified to suit his/her requirements.

An attempt is made to use a wide variety of different testing methods to decrease test method effects documented in literature on language testing (Bachman 1990). Also, most comprehension tasks are open-ended (short-answer questions and guided summaries in L1) giving the candidates some freedom of expression, but requiring them to construct meanings themselves instead of just recognizing the appropriate meanings, as in e.g. the multiple-choice technique. This approach is recommended by some test designers (e.g. Weir 1990). Because the number of candidates has been, and is likely to remain, relatively small, there is no compelling need to emphasize ease of scoring.

2.1. Structure of the Diploma test

The Diploma test consists of four components: a test of reading and writing, an oral interaction test, a language studio test, and a test of vocabulary and grammar.

The *reading and writing* test takes two hours. The reading part consists of three authentic texts and four tasks. It begins with a timed test of reading efficiency. The candidates are given 12 minutes to skim through a one- or two-page text and to find and construct answers to about five short answer questions. The task aims at measuring the readers' ability to read under a rather strict time

constraint, to locate specific information and to find the most important pieces of information in the text. Typically, the text includes a table or chart which also has to be understood.

After the first task-booklet is collected, the candidates are given the other reading and writing tasks, and they are free to allocate the rest of the time (1 h 50 min) as they wish. Recommendations for allocating time are printed on the cover of the test booklet.

The second reading task is a 100-130 word summary in L1 of a one-page text. The summary should give a reader unfamiliar with the original text a clear and accurate idea of the main content, and should show how well the candidate can understand the most important information given in the text. The assessors mark the summaries by checking whether certain pieces of content appear in the summary and whether they are linked clearly enough to show how they are related to each other.

The last reading passage has two tasks; first, the test takers answer half a dozen short answer questions focusing on the main points of the text. Second, they have to write in L1 the meaning of a number of underlined sentences in the text. Whereas the first part is intended to measure the comprehension of the main content of the text, the second aims at giving a picture of how detailed their comprehension is. The answers are scored in terms of how much of the factual content they convey; we try not to assess the style of the translations.

The above is the typical structure of the reading test, but we will probably use other test methods in the future. Sometimes the candidates can choose between two texts in the last part, if the background information indicates that the test takers' fields are sufficiently different. All reading comprehension tasks are double-assessed because they are relatively open-ended and thus somewhat subjective to mark. The same procedure is also used with listening comprehension tasks.

When giving the grade for reading comprehension the tasks focusing on how well the testee understands the main points of the text are weighted more than the reading efficiency task or the sentence translations. If the candidate has marked problems with reading rapidly, as indicated by a poor score in the first task or by inability to complete all tasks within the time limit, this will be reported in the certificate.

The writing section typically comprises two letter writing tasks, or a letter and a relatively short translation task into English, if the test taker has to translate a lot in his or her work. One of the letters is usually a complaint which requires the candidates to write about something not so pleasant and demand e.g. a compensation. The other task is typically a semi-formal letter concerning e.g. the arrangements of a visit to or from a foreign company or institution. The recommended time for the writing section is 50 minutes.

The writing tasks are assessed on a six-point holistic scale by two raters. If there is a mismatch in their ratings, they first discuss it together and then, if necessary, a third rater is consulted.

The second part of the Diploma test is what we call a *language studio* test that covers both speaking and listening skills, and takes about an hour and a quarter to complete. The *speaking section* is intended to tap those aspects of oral skills that are difficult to assess in our face-to-face interaction test. The most important of these is the ability to use language in different social contexts – a face-to-face test can reasonably be expected to show the candidate's skills in only one or two contexts. Also, the candidate's ability to construct and give a coherent and logical oral presentation is assessed in the lab test. The speaking test is similar to most other tape-mediated speaking tests in that the test takers hear the tasks from one cassette player and their responses are recorded onto another.

The studio test starts off by a few easy questions about the candidates and their work. These are not assessed but they serve the practical purpose of introducing the candidates in an easy way to the language lab and the tasks. Also, we get their names recorded on the cassettes, which ensures correct identification at the assessment stage. It should be mentioned that the test situation may be the first time ever in a language lab for some older people who went to school before the time of language studios. This is a problem which deserves further attention in the future, since some people's performance may suffer because of unfamiliar surroundings.

The first real task in the lab is reading aloud a short passage of text, e.g. a letter. Only the candidate's pronunciation is assessed in this part, which is written so that it includes several points found to be difficult for Finnish learners of English. The usefulness of this task has been discussed, since pronunciation is also assessed in the face-to-face test.

The second set of tasks covers a simulated telephone conversation and a number of different situations where the candidate has to react orally in an appropriate way. This is the most important part of the whole speaking section in the studio test because here we can put the candidates in a variety of different social contexts and see how they manage linguistically. The situations are sequenced so that they form two or three thematically connected plots. E.g. they may start with a following description:

> "You go to the airport to meet Mr/Ms Jones who represents one of your foreign clients. He/she holds a similar position in his/her company to yours. You have not met him/her before. Greet him/her and start small talk."

That is then followed by other episodes relating e.g. to showing the visitor around the company, entertaining him/her in the evening, and negotiating with him/her.

Candidates see and hear the descriptions of the situations in their mother tongue because we want to make sure that they understand the nature of the situations they are supposed to be in. If the instructions were in the target language, some of the problems the candidates have may be due to the fact that they have not understood properly the situation; this task is not aimed at measuring their comprehension but the ability to use appropriate language. Very often candidates' task is to initiate a discussion or to propose something. The only criterion used in assessing these mini-situations is the appropriateness of language for the particular purpose.

The speaking section in the lab finishes with a task where candidates have to prepare a short informal talk or presentation e.g. about their own country. They have five minutes to prepare and another five minutes to give the talk. A list of suggested topics is given to the candidates in the test booklet. The main criterion used to assess performance on this task is the ability to organize one's points into a coherent whole.

Everything that the candidate says is recorded on tape which is then assessed by two independent raters. As in other sections of the Diploma test, a third rater is consulted if the two raters cannot agree on a grade.

The second major part of the language lab test covers *listening comprehension*. The first task has typically been a partial dictation, which is argued to be a test of global comprehension (Johansson 1975). It is a form of testing where candidates are presented with a text which has a number of gaps. They listen to the 'text', usually twice, and fill in the gaps on the basis of what they have heard. Also, we believe, it can be used to test the understanding of specific points of what is heard, if the gaps are selected in certain ways.

The major part of the listening grade is based on performance in the second listening task. This is a rather long discussion or interview with one or more people. Candidates have to make notes and write short summaries of the points discussed in each part of the talk. In addition, they may have to answer a number of short answer questions focusing on the main content.

One of the rules in organizing the whole Diploma test is that the language studio test must always precede the face-to-face interaction. The idea is to give the candidates a chance to 'warm up' their oral skills before interacting with a real person; it may be that the candidates have not spoken English for some while before the test day, and we feel it is only fair to allow them some practice before the interaction test. Of course, it can be argued that the test takers may find the studio test hard if they have not practiced speaking for some time, but

since the greatest part of the speaking grade is based on performance in the interaction test, the above mentioned rule seems justifiable.

The face-to-face *oral interaction test* consists of two or three parts. For most candidates there are only two parts, although they are asked to be prepared also for the third one. The test starts with a formal presentation given by the candidate. Test takers are sent the task a few days prior to the test, so that they can prepare for it. Usually the task involves presenting the company or institution of the candidate, or telling about his/her work, or the products they make. The time allowed for the presentation is 10-12 minutes, during which the interlocutor usually interrupts the speaker with a couple of questions just to see how he/she copes with unexpected changes of topic. Candidates can use overhead projectors or other technical aids, but they are not allowed to read their presentations from paper.

The presentation is followed by an interview or discussion (depending of course on how the interaction flows) with the tester, who has taken notes during the talk to be able to link the discussion with what the candidate has presented. It should be mentioned that our testers receive some information about the candidates, their work, companies and language needs before the test, which enables them to plan some of the questions and topics in advance.

There is also a third part in the oral interaction test for those candidates whom the tester estimates not to be active enough during the discussion. This part is a so-called reversed-role interview where the candidate has to ask the tester some questions.

Originally we had the interview first and then the presentation at the pilot testing stage. That caused problems, however. The intention is that the presentation should elicit a sample of formal speech and the discussion a less formal, even informal sample. It turned out that if the interaction test started with the interview / discussion, it was very hard and unnatural for the candidates to revert to formal talk. In fact, we did not get practically any formal talk from some of the candidates. When the order of the two tasks was reversed, we could get a more balanced sample of both formal and informal registers.

There also was the reversed-role interview as a regular part of the test in the beginning. However, with so many proficient and interactive candidates, that part seemed out of place because it often interrupted an interesting discussion and imposed on the candidates an artificial task. Now it is only used with the more reticent candidates. Part of the problem with this task is the difficulty of finding plausible reasons why the candidate might want to ask questions from a foreign language teacher who often has limited experience of the business world of the candidates. Also, defining and operationalizing 'activeness' in interaction has been problematic. Asking questions is obviously typical of a person who is regarded as active in an interaction, but there is probably much more in the concept that is not so easy to turn into a test task, such as interrup-

ting, initiating topics, and demonstrating that one is interested in and reacts both verbally and non-verbally to what is said.

The oral interaction test is assessed by the interviewer immediately after the test. All performances are also double-assessed on video tape by a member of the testing group. If there are differing opinions, a third rater is consulted. The assessment is done both analytically and holistically by using a number of criteria which are specified later in this article. A summary of the criteria can be found in Appendix 1.

The last test of the examination is a test of *vocabulary and grammar*. These are tested separately for two reasons. First, they are at the core of language and their mastery is the basis for success in reading, writing, speaking and listening. Naturally, we recognize the fact that there is much more to language proficiency than vocabulary and grammar, but we believe that they are important enough to merit a separate test along with more skills-based tests. Second, we aim at giving a rather comprehensive assessment of the candidates' language proficiency, which is indicated by the number of different grades we report in the certificate. The mastery of vocabulary and grammar are there in the certificate mainly for the test takers' own benefit, so that they would also have feedback on these aspects of language.

Vocabulary and grammar are typically tested in a text context by using multiple-choice items or selective gap-filling techniques. When giving the grades for vocabulary and grammar, the candidates' performance on other tests is also taken into account, especially in borderline cases.

2.2. Assessment of oral skills

Developing a valid assessment scale for oral skills has been one of the key areas in the whole testing project. Oral skills are a very important part of the language proficiency needed in the work (e.g. Mehtäläinen and Takala 1990). However, there have been very few systematic attempts at testing oral proficiency in our country. For example, the largest examination in the country, the Matriculation examination, still lacks a speaking test. Except for schools of translation, the entrance examinations of the language departments at universities also lack tests of speaking.

Two approaches to the construction of the oral rating scales for the Diploma test were used. We surveyed the literature on language proficiency and oral skills in order to ensure that our criteria would cover all the relevant aspects of proficiency. We relied mostly on models based on the Canale and Swain tradition, such as that of Bachman (1990 and 1991), and more detailed analyses of oral skills (e.g. Bygate 1987). On the other hand, we surveyed several domestic and international assessment scales to see what kinds of approaches are most common. The construction and revision of the rating scales has probably been

the most time-consuming part of the whole project, but we feel that the effort has been worthwhile since designing the scale to a large extent equals defining the skill we attempt to test.

We now use an assessment scale that is divided into six ability levels and has subscales for six criteria plus a holistic scale. In addition, there are some minor criteria, which may be taken into account when giving the holistic assessment. The six main criteria are

Pronunciation
Speech Flow
Grammatical Accuracy
Vocabulary
Appropriacy
Discourse Management

The criteria are listed in the order they appear in the rating booklet. The order is based roughly on the relative difficulty and complexity of the criteria from the assessor's point of view, the first three being perhaps slightly more straightforward than the others. The small-sized rating booklet has a separate page for each criterion in an attempt to reduce the effect of a possible halo effect found out in some studies (e.g. Yorozuya and Oller 1980 and Lehman, this volume). Each page is divided into two parts: a six-point scale on which the rater marks the grade on that criterion, and space for the assessor's comments. The pronunciation scale is divided into two subscales, pronunciation of sounds and suprasegmental features (e.g. intonation), but the raters have to combine the two into one global grade for pronunciation.

We started with a clear-cut six-point scale, but soon it turned out that assessors wanted to be more specific and started putting their markings also on the borders of the levels. Now each of the six levels is defined into three sections: the rater can choose whether the candidate is e.g. a 'low', 'typical or mainstream', or 'high' four.

The six criteria are the basis for a global grade for oral interaction. The assessors are instructed to use the rough average of the analytic criteria when giving the overall grade, but they should not merely compute it mechanically. If it appears to the assessor that the candidate's performance on one or more of the criteria has a marked effect on the overall impression which one gets of the candidate's oral skills, then the assessor can give a lower or a higher grade than the mere average would indicate. In these rather rare cases the assessor has to write down reasons for deviating from the common procedure, and these reasons have to be approved by the second rater.

In addition, there are some other features besides the six criteria which are allowed to influence the global grade. These are listed in the rating booklet just before the scale where the global grade is marked. The features are:

Textual organization
Dependence on the interlocutor for comprehension (clarification, repetition)
Relevance and adequacy of content
Use of strategies to get around difficulties in communication and to enhance performance
L1 interference
Interactiveness

Each feature is provided with a 'scale' in the form of a line with the low and high ends marked. The assessor is instructed to rate candidates simply as 'low', 'average' or 'high'. An example is provided below:

Textual organization

I----------I
poor good

Some of the features overlap to some extent with the six criteria, but since they can affect a person's communicative efficiency in a foreign language we wanted to include them separately in the rating booklet. We attempted to use six-point assessment on some of these minor criteria (e.g. textual organization, use of strategies), but that turned out to be rather unreliable, so a more simple way of rating is used instead. In addition, some criteria are not needed to assess most candidates (e.g. relevance of content, use of strategies), but occasionally they might be useful and give some additional information about the candidate's oral skills. Usually, the additional features do not have any effect on the overall oral grade, but in some cases, often in the borderline cases, they play some part.

Contrary to many other rating scales, we have decided not to use the term 'fluency'. Instead, we have adopted the expression 'speech flow', because it seemed to us that fluency is one of the fuzziest and most ambiguous terms in applied linguistics (see Lehtonen et al. 1977; Sajavaara and Lehtonen 1978; Richards et al. 1985). According to some definitions, fluency almost equals language proficiency, whereas sometimes a far more restricted views are expressed. It also appears that the criterion called 'fluency' has different meanings in the international tests we have looked at. We attempted to use a very narrow definition that concentrates on hesitation and unnatural pauses (caused by the need to search for expressions).

2.3. Certificate

Certificates are issued within two weeks but preliminary results of the test can be obtained earlier. The certificate presents the candidate's language profile: it describes his or her command of several aspects of the foreign language. Such

a detailed certificate can also provide diagnostic guidance for future learning. More generally it is hoped that the examination will encourage and stimulate language learners and provide a beneficial 'washback' effect on language teaching and learning in the professional sector.

The aspects of language that are reported in the certificate are as follows:

Overall grade

Speaking
Listening
Reading
Writing

Speech Flow
Appropriacy
Grammar
Vocabulary

Some of these are also described in writing (see Appendix 2). The backside of the certificate contains a short description of the content of test and the 6-point overall proficiency scale.

Some of the points in the certificate probably require some clarification. The overall grade is the rough mean of the four major skills. If e.g. two skills have the grade 4 and two have 5, the productive aspects play a decisive role. Thus, if speaking and writing grades are lower than the comprehension grades, the candidate will get a lower overall grade.

Appropriacy is mostly decided on the basis of a specific test that takes place in the lab, but the estimates of appropriacy coming from the oral interaction test and the writing tasks are also considered. Grammar and vocabulary grades are similar: they are also based on specific tests of grammar and vocabulary but information from the productive tests is used to support assessors' decisions.

2.4. Reliability of ratings

As described above all tasks are rated by at least two assessors, including the comprehension questions. Only the more mechanical vocabulary and grammar tests are usually scored only by one person who uses a scoring key compiled by the testing group. Occasionally another member of the group is consulted if an unforeseen answer appears in the gap-filling tasks.

Since the number of the test takers has been relatively small and the performances have been assessed by the same small group of assessors who consult each others if the ratings do not match, there has not yet been a compelling need for

more extensive studies on the reliability of assessment. However, we have attempted to ensure reliability and validity of the oral grades by providing training for our interviewer-raters, since the oral interaction test is the only part of the Diploma test where we partly rely on raters outside our small circle of test developers. We are also prepared for larger numbers of test takers and assessors by compiling a benchmark video to standardize the rating in oral tests. Also, we have carried out some limited reliability checks on the assessment of oral skills, which will be reported later in this article.

Our oral assessors / interviewers are language teachers, most of whom are native speakers of English (or German or Swedish, depending on the language tested). They give the first assessment of the candidates' performance right after the test; these ratings are then checked by a member of our testing group on video tape. We train the assessors during two one-day training sessions. The first day typically includes more general lectures and discussions on assessment and interviewing techniques and on the particular assessment system used in the Diploma test. We have also used practice interviews that are viewed and commented by other assessors. During the second training day the new assessors discuss the candidates they have assessed on video at home. Some time ago we distributed the landmark video to the acting interviewers to help them in their work and to improve reliability. After that, however, the number of the candidates taking the test declined further as a reflection of the economic depression in our country, so that we have not been able to study how much the new video has improved rating.

A small scale inter- and intra-rater reliability study has been done on the oral interaction test performances. The ratings done by the two most active members (in terms of the number of ratings done) of the testing group. The intercorrelation between the two assessors' overall ratings was .79 (p=.000; n=16). One of the raters did not use the landmark video to calibrate his ratings, which may explain the fact that the correlation is not higher. In addition, the rating sessions took place during a two-week period, so they were not very concentrated in terms of time. The following lists the intercorrelations on the analytic criteria:

Pronunciation	.72	p=.001
Speech Flow	.86	p=.000
Grammar	.75	p=.000
Vocabulary	.74	p=.001
Appropriacy	.68	p=.002
Discourse Management	.61	p=.006
Textual Organization	.60	p=.026

(n in Textual Organization = 11, in others n = 16)

The number of candidates in this study was small, but the coefficients clearly indicate that there was room for improvement. Fortunately, we have the possibility of meeting regularly and sorting out differences of opinion such as

reflected by the above figures which come from independent ratings prior to the standardization meetings. However, should the number of candidates rise dramatically, the system needs improvement. Part of it has already been done by the introduction of benchmark videos, as mentioned above. Also, the status of 'textual organization' has been lowered, since it could not always be assessed at all, and even in cases it could be done, a six-point scale appeared to be too detailed for the purpose. Now 'textual organization' (cohesion, coherence of presentation) is only an additional criterion, which may play part in some borderline cases.

The results also suggest that 'appropriacy' and 'discourse management' are more problematic to rate than the other criteria. Quite probably the definitions of the criteria need improvement and more concrete references to performance. It was perhaps surprising to notice that 'speech flow' was the most reliably rated aspect of oral skills, given the conceptual difficulty of defining 'fluency', from which it was derived. Apparently, we have succeeded in narrowing down the concept to a manageable whole from the point of view of rating – as far as the two raters studied are concerned.

Some intra-rater correlations were also computed. One of the assessors rated 18 candidates twice with a roughly two-month interval between them. The correlation between the two ratings of overall oral proficiency was satisfactory (.89, p=.000) given the fact that it was done without the help of a benchmark video. There is some indication that the benchmark video improves intra-rater reliability: one of our raters achieved almost perfect agreement with himself on four randomly selected candidates when he used the benchmarks to calibrate ratings (.97, p=.014).

2.5. Validity of the Diploma test

We attempt to achieve validity by designing the tests to cover all the relevant aspects of language skills and selecting tasks and texts that are appropriate to the candidates.

The tests cover all four traditional aspects of proficiency – speaking, listening, reading, writing – instead of just the written skills, which has been the case in so many tests in our country in the past. An attempt is done to include all aspects of language proficiency included in modern theories of language skills (e.g. Canale and Swain 1980, Faerch et al. 1984, Bachman 1990, Bialystok 1991). The need to have several different tests which cover different aspects of language is also emphasized by empirical findings that suggest that language competence is divisible (reviewed in Skehan 1988).

More specifically, we have adopted many ideas of the so-called communicative testing, for example, an emphasis on the use of language for particular purposes – rather than just knowledge of language – and the use of authentic texts and

tasks (Carroll 1980; Weir 1990). We also attempt to test the sociolinguistic aspects of language, i.e. the appropriate use of language in various settings.

To ensure valid content for the tests we have carried out extensive surveys of the needs of foreign languages in various areas of work (Mehtäläinen 1987a, 1987b, 1988, 1989; Mehtäläinen and Takala 1990). They form the basis for the selection of task types for the tests. In addition, we obtain a more detailed, individual information about the candidates who enter the examination through a fairly detailed questionnaire (see Appendix 3). The interviewer of the oral interaction test can base his/her questions partly on this information. Also, some of the tasks may be changed, if the candidate's needs are very special.

The examination was pretested on two occasions. The feasibility of many of the task types, but only a few of the actual tasks that were later used in the real test, were pretested in a placement test for a language course at a big Finnish industrial company. On the basis of the trial some of the task types were discarded while others were accepted as potentially suitable. Especially valuable was the experience we got in organizing an oral test and doing the interviewing and assessing. We also found out how laborious and time-consuming a really tailor-made test construction can be, since some of the tasks had as many as ten different versions depending on the position of the person in the company. Accordingly, a less tailor-made approach was adopted for the Diploma test, although there is some attempt left to modify certain tasks even at the level of individual test takers.

The second pretest was carried out when the Diploma test was introduced to the potential customers. About thirty of them were invited to take the test for free. The results were analysed and some changes were made to the tasks and the organization of the tests. The most notable finding was that we had to reorder the oral interaction test: as explained in more detail elsewhere in this article, we changed the places of the oral presentation and the interview. Another change worth mentioning was the removal of the simulations of various language use situations from the written part. It appeared that the candidates found it rather odd that they should write down what they would *say* in this or that situation. Also, written answers do not necessarily show what the test takers would say under the real time pressures of oral communication. Since the pretest, these simulations have taken place in the language studio, where the candidates have to react orally.

The feedback we gather by post-test questionnaires has been overwhelmingly positive, which goes some way in demonstrating empirically that we have a fairly valid test. Most candidates appear to believe that the test was fair and measured their ability to use the foreign language. The most common comments have been that the Diploma test is the most comprehensive and thorough test they have ever taken, and that they were satisfied with the fact that they had the opportunity to really use the language in the test, not just answer questions on grammar. There is also some negative feedback on certain problems with

administering the test (especially on the lab test because the candidates can hear each other talking which may disturb some), or, occasionally, on the unsuitability of some tasks for certain candidates.

We have also computed the correlations between the grades that are reported in the certificate and the candidates' assessment of their own English language skills which are collected through the pre-test needs analysis questionnaire. The correlations are all in the range of .4 - .5 (p=.000; n=68) on the aspects that are assessed (overall success in FL communication, appropriacy, grammar, vocabulary and pronunciation/spelling). It has been shown that self-assessment can be fairly reliable as the studies reviewed by Oscarson (1989) suggest. Thus, it is encouraging to find out a significant, even if not a very strong, relationship between the candidates' own estimates and the grades awarded on the basis of our tests.

2.6. Future

The whole testing system is under revision at the moment. The old six-point assessment scales will be replaced by a new nine-point scale to allow a more detailed assessment at the lower levels, since there are two new tests under construction for basic and intermediate proficiency levels. The overall proficiency scale is roughly comparable to the English Speaking Union's nine-point scale, with the exception that our level 1 is close to their level 2, because we feel that the ESU's level 1 is too low to be worth testing. The new basic test covers levels 1-3 on our scale, the intermediate covers levels 3-5, and the old Diploma test levels 6-8. The basic and intermediate tests are more general in their topics than the Diploma test.

REFERENCES

Bachman, L. 1990. *Fundamental Considerations in Language Testing.* Oxford University Press.

Bachman, L. 1991. What Does Language Testing Have to Offer? *TESOL Quarterly* 25 (4), 671–704.

Bialystok, E. 1991. Metalinguistic Dimensions of Bilingual Language Proficiency. In E. Bialystok (ed.) *Language Processing in Bilingual Children.* Cambridge University Press, 113–140.

Bygate, M. 1987. *Speaking.* Oxford University Press.

Carroll, B. 1980. *Testing Communicative Performance.* Oxford: Pergamon.

Faerch, C., K. Haastrup, and R. Phillipson 1984. *Learner Language and Language Learning.* Copenhagen: Gyldendals Sprogbibliotek.

Johansson, S. 1975. Partial Dictation as a Test of Foreign Language Proficiency. In S. Johansson, *Papers in Contrastive Linguistics and Language Testing.* Lund Studies in English 50. Lund: CWK Gleerup, 123–149.

Konttinen, R. *Research into the Role of Final Examinations in Secondary Education.* Paper presented at the Council for Cultural Co-operation Workshop, Jyväskylä, June 15-18, 1993.

Kuru, I. 1993. *Exercises as the Indicator of Didactic Thinking in the Upper Secondary School Swedish Instruction.* Unpublished Licenciate Thesis. Department of Education, University of Jyväskylä. (In Finnish)

Kärkkäinen, K. & Takala, S. 1988. Developing the language tests of the matriculation examination (In Finnish). In V. Brunell & J. Välijärvi (Eds.) *Yleissivistävä koulutus: Näkymiä ja tulevaisuuden haasteita.* Institute for Educational Research, Publication series B, Nr. 25, 159-189.

Lehtonen, J., K. Sajavaara, and A. May 1977. *Spoken English. The Perception and Production of English on a Finnish-English Contrastive Basis.* Jyväskylä: Gummerus.

Mehtäläinen, J. 1987a. *Teollisuuden toimihenkilöiden kielitaidon tarvetutkimus (The Need of Foreing Language Proficiency in Industry).* Teollisuuden koulutusvaliokunta (Training Commission for Industry): Publication 8.

Mehtäläinen, J. 1987b. *Helsingin kaupungin henkilöstön kielitaidon tarve (The Need of Foreign Language Proficiency of the Employees of the City of Helsinki).* Statistical centre of the City of Helsinki, Report 8.

Mehtäläinen, J. 1988. *Suomen Kunnallisliittoon kuuluvien kuntien viran- ja toimenhaltijoiden kielitaidon tarve (The Need of Foreign Language Proficiency of the Employees of the Finnish Municipalities).* Centre for Language Learning and Teaching, University of Jyväskylä. Manuscript.

Mehtäläinen, J. 1989. *Kaupan henkilöstön kielitaidon tarve (The Need of Foreign Language Proficiency in Trade).* Centre for Language Learning and Teaching, University of Jyväskylä. Manuscript.

Mehtäläinen, J. and S. Takala 1990. *The Need of Foreign Language Proficiency in Trade and Industry in Finland.* Manuscript.

Oscarson, M. 1989. Self-Assessment of Language Proficiency: Rationale and Applications. *Language Testing* 6 (1) 1–13.

Richards, J., J. Platt, and H. Weber 1985. *Longman Dictionary of Applied Linguistics*. Longman.

Sajavaara, K. and J. Lehtonen 1978. Spoken Language and the Concept of Fluency. *Language Centre News* 1/1978 (Special Issue: Focus on Spoken Language), 23–57.

Sajavaara, K. 1992. Designing Tests to Match the Needs of the Workplace. In E. Shohamy, A. R. Walton & Morfitt, C. A. (eds.) *Language Assessment for Feedback: Testing and Other Strategies* (122-144). National Foreign Language Center. Dubuque, IO.: Kendall/Hunt Publishing Company.

Skehan, P. 1988. Language Testing. *Language Teaching* 21 (4), 211–221.

Takala, S. (Ed., 1993). *Suullinen kielitaito ja sen arviointi [Oral proficiency and its assessment]*. University of Jyväskylä. Institute for Educational Research. Publication Series B: Theory into Practice Nr 77 (269 pp.).

Weir, C. 1990. *Communicative Language Testing*. New York: Prentice Hall.

Yorozuya, R. and J. Oller 1980. Oral Proficiency Scales: Construct Validity and the Halo Effect. *Language Learning* 30, 135–153.

Appendix 1

Oral assessment criteria used in the Diploma test: Key points

Key points and concepts

ANALYTIC CRITERIA

PRONUNCIATION

- pronunciation of individual sounds
- word-stress
- suprasegmental phenomena: intonation, pitch, rhythm

SPEECH FLOW

- unnatural pauses, hesitation
- appropriate rate of speech
- ease vs. struggle

GRAMMATICAL ACCURACY

- systematic vs. random errors
- basic vs. complex structures & errors

VOCABULARY

- precision, adequacy and range

APPROPRIACY

- adjustment of speech to suit situations; use of suitable register in terms of role and participants
- showing awareness of cultural differences
- level of formality
- level of directness (eg forms of address and politeness)

DISCOURSE MANAGEMENT

- understanding and reaction; backchannel signals
- turn-taking
- starting, maintaining, and closing discourse

COMMUNICATIVE EFFECTIVENESS

(OVERALL GRADE)

- holistic assessment / outcome/ goal achievement on the basis of all the analytic criteria and the following points:
- dependence on the interlocutor for understanding (clarification, repetition)
- relevance and adequacy of content
- use of strategies to get around difficulties in communication and to enhance performance
- L1 interference
- interactiveness (withdrawn vs. interactive)

Appendix 2

Sample Certificate of the Diploma test (translated from Finnish)

THE FINNISH FOREIGN LANGUAGE DIPLOMA FOR PROFESSIONAL PURPOSES

Finn Sisu, Jr. (098765-4321) has taken the test intended for middle management in business, on December 18, 1990. On the basis of a battery of oral and written tests, his proficiency in English has been assessed on a scale of 1-6 (see overleaf) as follows:

Overall proficiency 3

* Comprehension of spoken English	4	* Pronunciation	4
* Speaking proficiency	3	* Fluency of speech	3
* Comprehension of written English	4	* Appropriacy	3
* Writing proficiency	3	* Grammar	4
		* Vocabulary	4

This person has achieved an **intermediate** level of language proficiency in English. He understands spoken English well and can present what he intends to say quite clearly but with some effort. His language use is sufficiently active. He can read texts that deal with general topics and with his own field without much difficulty and understand all main ideas. In writing he makes mistakes which may impair comprehensibility and require some effort from the reader. His vocabulary is fairly large and sufficient for most purposes.

On this level of proficiency a person can be expected to be able to participate in general tasks and situations that require a command of English.

The test has been prepared and the performance assessed by the University of Jyväskylä and it has been taken at a testing session arranged by the Institute of Marketing.

Jyväskylä, January 25, 1991

Heikki Hirvinen Chairman of the Diploma Board	Sauli Takala University of Jyväskylä	Matti Honni Institute of Marketing

The holder of this certificate has taken a 5-hour battery of tests, which measure the ability to understand and use oral and written English. Performance has been assessed in terms of communicative efficiency and appropriacy, linguistic accuracy, comprehensibility, fluency, pronunciation and spelling.

An examinee is required to:

1. tell orally about him/herself, his/her work, company/organization and home country; discuss matters relating to work, and cope with various communication tasks and situations which relate to work and hosting foreign visitors;

2. listen to and understand the main content of discussions, interviews, or reports concerning his/her own field; the listening comprehension tasks require the understanding of speech presented at varying speeds and in varying accents;

3. read authentic texts dealing with his/her own field, and demonstrate both an overall and detailed understanding of the texts;

4. write official and semi-official letters or other documents or translate short texts relating to his/her work.

Task performance scale

Level 6 **Excellent**
Communicates efficiently and actively in demanding tasks and situations. Approaches native-speaker proficiency.

Level 5 **Very good**
Communicates actively and appropriately in demanding tasks and situations. Has a good command of language and makes few mistakes. Some evidence of mother-tongue usage.

Level 4 **Good**
Can handle tasks and situations that are familiar. Can take a fairly active part in discussion. Mistakes occur but they rarely cause problems of comprehension. Mother-tongue usage quite recognizable, but not disturbing.

Level 3 **Intermediate**
Can handle familiar tasks and situations fairly well, but prefers to respond rather than initiate. Passive knowledge of vocabulary clearly large than the active vocabulary. Limitations of language proficiency cause communication to be occasionally non-fluent and lead to some misunderstanding. Obvious accent and mother-tongue usages.

Level 2 **Marginal**
Can handle routine tasks and situations without too much effort. Knows enough grammar and vocabulary to be able to get the gist of most commonly o ring situations. Limited proficiency can cause some problems of comprehens nd misunderstandings.

Level 1 **Basic**
Can cope with only the simple routine tasks and situations. Has only mastered some basic element of the language. Deficiencies in proficiency cause frequent comprehension probl ns and can lead to misunderstandings and breaks in communication.

Appendix 3

Sample questions from the questionnaire that the Diploma testees fill in prior to taking the tests

Question 10.

A. Please, estimate how often you need English in your work in the following situations. Mark the number describing the frequency of each situation in the box.

0 never
1 a few times a year
2 monthly
3 weekly
4 daily

B. Please, estimate the adequacy of your language proficiency in handling the situations in which you have needed English in your work. Mark the number describing the adequacy in the box.

1 completely inadequate
2 fairly inadequate
3 almost adequate
4 completely adequate

C. Please, estimate how important the situations below are for you in your work now or in the future. Mark the number describing the importance in the box.

1 not at all important
2 not very important
3 important
4 very important

Example: If you need English in customer service weekly, and consider your language proficiency almost adequate for that language use situation, and consider the situation very important for your work, your answer will look like the following:

A B C
3 3 4

I. ORAL SITUATIONS	A	B	C
a. customer service, guidance			
b. hosting visitors, social use of language, small talk			
c. interpreting at work			
d. interpreting in social situations, when other things than work are discussed			
e. taking messages, answering inquiries on the telephone			
f. initiating client contacts, conducting demanding discussions on the telephone			
g. visiting fairs / visiting foreign companies			
h. managing a trade fair stand			
i. attending training sessions			
j. customer training / job instructions			
k. negotiations and meetings			
l. restaurant and travelling situations			
m. presenting one's own company / work			
n. other situations, what?			

...

II. WRITTEN SITUATIONS	A	B	C
a. reading brochures, instructions for use, service and maintenance instructions, shipping documents, etc			
b. reading telegrams, telex and telefax messages and business letters			
c. reading professional journals and professional literature			
d. reading reports, agreements, memoranda, minutes of meetings			
e. translating from English into Finnish			
f. translating from Finnish into English			
g. filling in forms			
h. writing brochures, instructions for use, service and maintenance instructions, transportation documents, etc			
i. writing telegrams, telex and telefax messages and business letters			
j. writing reports, agreements, memoranda and minutes of meetings			
k. situations related to use of computer			
l. other situations, what?			

11. If you translate in your work, what kind of texts do you translate?

...

15. Please, estimate your mastery of English in the following aspects of language. Mark the number describing your estimate in the box.

1 poor
2 fair
3 satisfactory
4 good
5 very good
6 excellent

ORAL		WRITTEN	
Pronunciation		Spelling	
Grammatical accuracy		Grammatical accuracy	
Vocabulary		Vocabulary	
Appropriate use of language in the situation		Ability to write in the appropriate style	
How well do you think you generally succeed in communication?		How well do you think you generally succeed in communication?	

TEST FOR MEASURING THE UNDERSTANDING OF ICONICAL TROPES

Christo Kaftandiiev and Felly Stoyanova
Sofia University

The article deals with the role that iconical tropes play in understanding and argues that their importance is often neglected in testing at the expense of measuring verbal understanding. An experimental test of trope comprehension is described together with the results of its trial run.

INTRODUCTION

Tropes[1] play a very important role in semiotic processes. The processes of tropisation (coining of tropes) determine the functioning of signs both in a diachronic and in a synchronic aspect. There are different reasons for this and one basic reason is determined by the associative nature of human thinking. A possible explanation of this fact is that the objects and phenomena are not clearly and definitely separated from one another in time and space (respectively, their signs). These characteristics of human thinking and of a sign system combined with such semiotic phenomena like the law of language economy, the tendency towards sign expressiveness, unification and differentiation determine the key place of tropes in the processes of semiotic evolution and human communication.

These and other reasons make some researchers believe that there are only two ways for the development of signs - tropisation and desemantisation (Gak, 1974). Other researchers emphasize that tropes are "the core of linguistic (and especially poetic) creativity". (Crystal, 1988)

The characteristics of associative thinking and the processes of linguistic signs' tropisation determined by it find their adequate realisation on different language levels. From a diachronic point of view the process of tropisation develops with different intensity depending on a number of extralinguistic characteristics. This phenomenon find confirmation in the works of some leading anthropologists like Fraser, Levi-Strauss, and others. The same applies to the formal sign characteristics too with regard to the sign system evolution.

[1] "Trope" ('turn', 'conversion') is an instance of figurative language in which words are used in a way that effects a conspicuous change in their standard meaning.

Tropisation is the core of hierographic alphabets mostly because they are built on the basis of associative signs. The associative sign is a kind of picture of a concrete object which corresponds to the word and is connected with the object through association. For example, a picture of the sun could correspond to the word "day" (Gelb, 1963).

Despite the importance of associative thinking and the language mechanisms which help its functioning, tests for assessing the understanding of tropes that are expressed through different sign systems, do not possess a role that is congruent with their real place in the process of a sign system functioning.

The reasons for this can be related both to the test makers and to the authors of syllabuses. The main reason can be explained with the nature of an institution that make the syllabuses. Most often these syllabuses are directed towards the verbal sign system, and as far as the learning of the mother tongue is concerned, towards grammar. One of the chief reasons for this obvious lack of correspondence with the real needs is the "enormous power of institutional inertia" (Spolsky, 1991). The basic problem at the moment lies not that much in the test validity but in the degree of the syllabus validity. These conclusions are supported in a number of scientific articles.

> "How good the syllabus is, is a problem for syllabus designers, materials writers and so on ... But the validity of the syllabus itself is not essentially the concern of the achievement tester. The validity of the syllabus must be taken for granted." (Alderson, 1988)

The absurdity of this situation with the syllabuses is more than obvious especially when bearing in mind that no single test can be considered valid without the examination of its validity. It is quite natural that the high degree of intuitiveness applied to the syllabus design will reflect on the test design too.

> "... the language testers, as applied linguists, must respond to the practical need for more appropriate measures of language abilities for use in language acquisition and language attrition research, language program evaluation, and for making decisions about individuals' attained levels of competency with respect to various educational and employment requirements. I believe that most language tests currently available are inappropriate for these purposes because they are based on a model of language ability that does not include the full range of abilities ...
> Language tests are based on either course syllabi or theories of language proficiency, both of which represent the subjective judgement of the individuals who devise them, and tests are therefore neither infallible nor exhaustive definition of the abilities we may want to test." (Bachman, 1990).

These tendencies regarding the syllabuses design lead to a natural violation of the test validity because they do not measure sufficiently well one of the key linguistic phenomenon - the tropes, notwithstanding the fact that in some studies (Bachman, 1990) understanding of tropes is defined as an important characteristic of the sociolinguistic competence.

> "The final aspect of sociolinguistic competence to be dealt with here is that which allows us to use and interpret cultural references and figures of speech. Many of these will be incorporated, with set meanings, into the lexicon of any language, and can thus be considered part of lexical, or vocabulary, competence. Nevertheless, knowledge of the extended meanings given by a specific culture to particular events, places, institutions, or people is required whenever these meanings are referred to in language use. For example, to interpret the following exchange, the language user would have to know that 'Waterloo' is used linguistically to symbolize a major and final defeat with awful consequences for the defeated ...
> Although individuals from different cultural backgrounds will, no doubt, be able to attach meaning to figures of speech, the conventions governing the use of figurative language, as well as the specific meanings and images that are evoked are deeply rooted in the culture of a given society or speech community, which is why I have included them as part of sociolinguistic competence."

One of the main purposes of this research is to create a test battery for measuring the understanding of tropes through different sign systems. Only part of it, the creation of a test measuring iconical tropes understanding, is described in this paper. The orientation towards an *iconical* sign system is not accidental since the most important part of current language at school is directed towards the *verbal* sign system and above all towards its official variant.

The neglecting of the iconical sign system in the syllabuses as well as in tests can lead to some level of illiteracy bearing in mind the importance of iconical system in the end of the 20th century. Support for this view can be found in the studies of leading theoreticians of mass media communication. Marshal Mcluhan developed the idea of the balance of feelings that was seriously disturbed after the creation of the phonetic and linear alphabet and then again developed by electronic medias. Today communication is determined to a great extent by the most powerful media - the television. Under the influence of electronic medias people develop a stereotype for receiving information through iconical signs; the participant in communication constantly looks for this type of sign, perhaps in most cases unconsciously. This is why, the concepts of some semioticians are of such current importance.

> "The concurrence of sound and visual systems is characteristic of contemporary society ... It is quite possible that a small part of truth is contained in the numerous appeals to take heed of the increasing role of visual systems in contemporary society ...
> We can formulate the question: Would not the boy, we spoke about earlier, find himself in the position of a man who is forced to replace again the sound language with a visual one as thousands of years ago the visual language was replaced by the sound one?" (Ivanov, 1978)

1. METHOD

1.1. Description of Iconical Tropes Paradigms

It is possible to create different iconical tropes paradigms and, on their basis, to develop the respective test battery. One of these paradigms could be created on the basis of different types of images - photographs, paintings, cartoons, comics, maps, graphs, schemes, diagrams. This paradigm could be further developed according to the continuum concrete-abstract trope images as in the image classification done by Abraham Moles (Moles, 1981). These types of paradigms are very important because they cover all types of images and especially the more widely used, for example, the images published in magazines like Times and Newsweek. The images used in these magazines are models of high quality iconical interpretation of information. A very interesting semiotic law of the probability degree of trope interpretation of images could be formulated studying these magazines: The more abstract the image used in the press is, the higher the probability of its trope interpretation. We could formulate such a hypothesis or law with reference to advertising images as well but in this case the level of concreteness or abstractness will not be of importance. This law could be formulated approximately in this way: The probability for an advertising image to be interpreted as a trope is higher than a similar interpretation of an image used in other mass media genres.

Advertisement trope images are very suitable for test items. The value of these images is determined by the much higher level of intensity of communication processes in advertising compared to the processes in other types of mass communication. This strategy is supported, for instance, by the way in which Dutch researchers selected texts for test items measuring the extent of knowledge of a foreign language.

> "Tests of reading have been developed for French, German and English. They consist of about 25 open and closed questions about a number of authentic texts like leaflets, news items, announcements, advertisements, etc." (Wijgh, this volume)

These researchers defined 19 specific objectives concerning second language learning, categorized in four domains, two of which are the compensatory strategies and techniques and the sociocultural competence. The test on the understanding of trope iconical figures is directly connected with these two domains.

The test can be developed on the basis of other characteristics depending on the type of iconical trope in the continuum homogeneity - heterogeneity — metaphors, metonymies, hyperbole, litotes, symbols — or on the iconical trope types in the continuum aiconicity - high level of iconicity (Lossev, 1982).

1.2. Description of the Test

The test presented in Figures 1 and 2 is based on two models which are interpreted in two three-dimensional matrices. The different iconical tropes, like the different kinds of images through which they are realized, are the two basic elements in the matrices. Basic elements of the matrices are the different iconical tropes. The test items that contain metaphors and metonymies are the most frequent ones because of their role and quantity in the communication processes. The iconical metaphor could be defined as a figure in which the images of objects, activities and phenomena are transferred to the images of other objects, activities and phenomena on the basis of some significant element actualized by the context. The classification of iconical metaphors can be done in the same way as that of verbal metaphors. That is why, a new classification is proposed:

a) metaphors, in which the images of objects are transferred to the images of other objects;
b) metaphors, in which the images of objects are transferred to the images of activities and phenomena, and vice versa;
c) metaphors, in which the images of activities and phenomena are transferred to other images of activities and phenomena.

There are examples of these types of iconical metaphors in the test.

Metonymy is the second, actively used trope in the test. The metonymy can be defined as a figure in which the images of objects, activities and phenomena replace the images of other objects, activities and phenomena on the basis of some significant element, actualized by the context. The classification could be done in the same way as in the previous one. Iconical synechdoches, hyperbolae and litotes are also used in the test. One of the most important characteristic of the test is that the images must be accompanied by part of the advertisement's verbal text. It is determined by one of the basic differences between the verbal and the iconical signs — the much more clearly expressed polysemy of the latter. The role of verbal signs here is to fix the attention to one of these

numerous significances, or according to Barthes (1967) to "anchor" these numerous significances in order to put an end to the terror of uncertain signs. All these images are built by numerous significances (much more than verbal signs are), and for this reason, it is quite natural to have more interpretations. One of the main objectives of the verbal text in advertising (the headline, the copy and the tagline) is to anchor the consciousness to one of these significances. The same is also true of the other iconical genres of mass communication and even of art images.

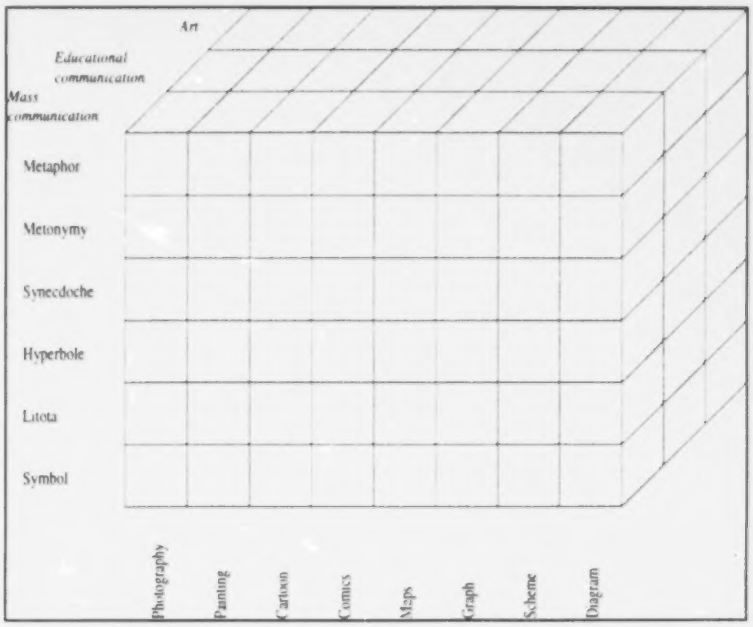


Figure 1. Classification of Iconical Tropes - Part I.

The way a sign functions in advertising is very interesting from a semiotic point of view — the integration of the two basic sign systems is very strong and it is quite reasonable to put the question: Could we speak of different sign systems or of a new sign system designed out of the verbal and iconical subsigns? For this reason, the headlines are given in the beginning of the test (the most important verbal part of the ad and this part of the copy in which the trope characteristics of the image are interpreted).

The third important part of the matrix (Figure 2) is connected with the behaviour objectives of learning. Bloom's taxonomy is used in its variant adapted for second language assessment. This test measured only part of the cognitive behaviours - the ability to analyse utterances or patterns in terms of implicit meaning.

> "Most of the items in this area are still experimental, for only recently have the implicit and implicate meanings of utterances and cultural patterns received much attention. It must be remembered that these types of behaviours are usually achieved only by advanced students, and that many advanced students develop these behaviours on their own after years of experience with the language and culture." (Valette, 1971)

Another paradigm may be used in this test as well, for example, from an advertising point of view, depending on what is advertised - products, services, trademarks, institutions, the product's life cycle, the appeals used and so on. This paradigm, however, will be more suitable for an article in a scientific journal like Journal of Advertising Research.

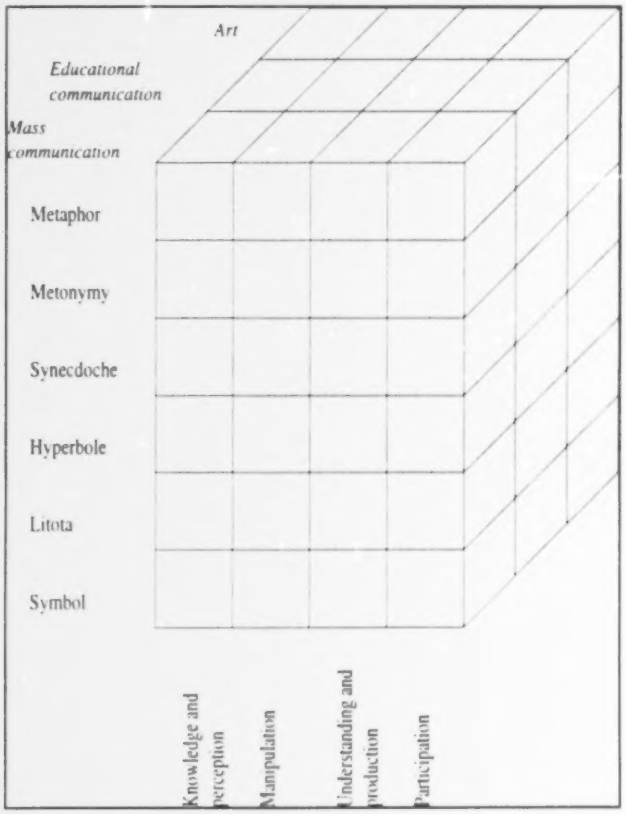


Figure 2. Classification of Iconical Tropes - Part II.

2. RESULTS

2.1. Analysis of the Tryout Data

2.1.1. Item Parameter

The item difficulty level P_g (the proportion of examinees getting the item right) varies between 0.18 and 0.85., and the mean (p) of p_1, p_2 ... p_{25} is 0.59 (p = 0.59). The mean of the item difficulty levels is relatively good for this kind of test (multiple choice items with 4 alternatives).

The item discrimination is also relatively good:

- The biserial correlation (r_{+bis}) of success on the items to the score varies from +0.23 to +0.87, and these values are significantly greater than zero beyond the 0.05 level. The mean of r_{+bis} is 0.50.
- The biserial correlation (r_{-bis}) of the single distractors to the total score is always ≤ 0, and 79% of these coefficients are significantly greater than zero beyond the 0.05 level.

In the item analysis we also used the three-dimensional normal ogive model (Lord & Novick, 1968). The three item parameters are: a_g - the discriminating power of item g; b_g - the index of the difficulty level of item g; c_g - the probability of guessing correctly. In this test c_g = 0.25 for g = 1, 2, 3, ... 25.

The item characteristic curve of all items was built and analysed on the basis of this model. The item characteristic curves of some items are shown in Figure 3.

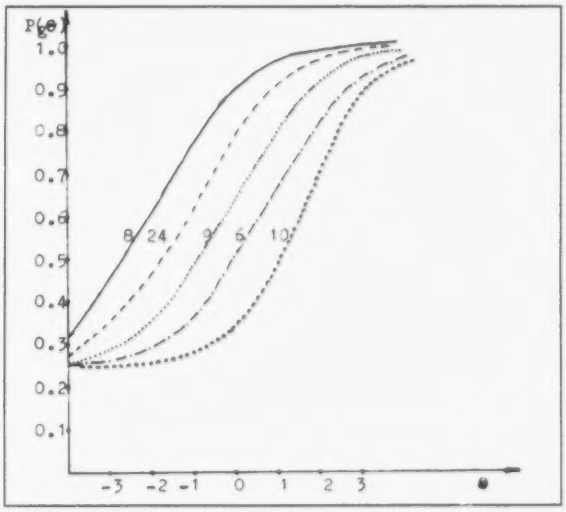


Figure 3. Item Characteristic Curves for Some Test Items.

The mean of the values of a_g (g = 1, ... 25) is 0.67. This value is sufficient for our purposes because in this way the test items provide a possibility for better discrimination in a larger interval of the latent continuum.

The mean of the values of b_g (g = 1, ... 25) is -0.47 and for the different items this parameter varies between -2.95 and 1.86, that is, the test items with their difficulty level cover well the interval (-3; +2) of θ.

The information function I_g (θ) of the test items was also calculated. In Figure 4 are shown the item information functions of the test items (No. 6, 8, 9, 10, 24) whose item characteristic curves are shown in Figure 3. As can be seen, the maximum of the information function is attained at different value of the latent attribute.

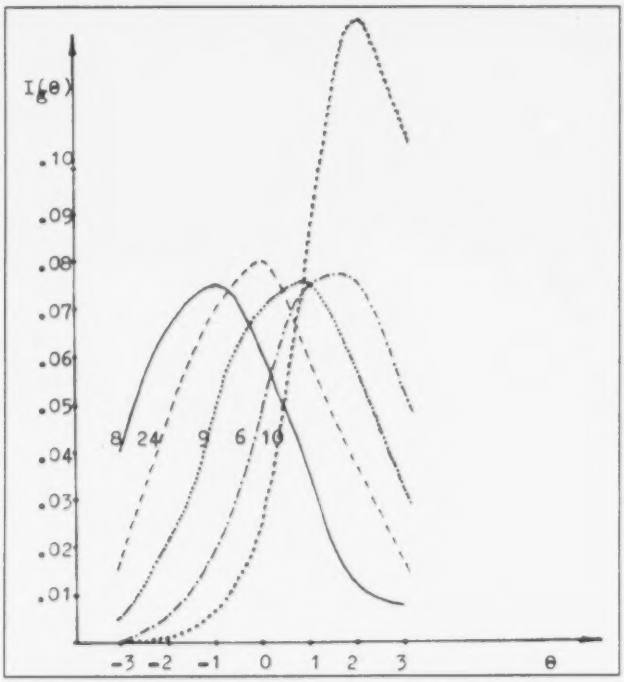


Figure 4. Information Function for Some Test Items.

In Table 1 (next page) the values of the item parameters (p_g, $r_{.bis}$, $r_{.bis}$, a_g, b_g) for all test items are given.

Table 1. Item Parameters for Test Items

Item no.	P_g	A	B	C	D	b_g	a_g
1	0.72	-0.46	-0.52	+0.56	-0.06	-1.04	0.67
2	0.40	-0.51	-0.26	-0.15	+0.34	0.76	0.36
3	0.31	-0.25	+0.31	-0.42	-0.02	1.63	0.32
4	0.82	-0.39	-0.83	+0.87	-0.62	-1.07	1.76
5	0.50	-0.88	+0.31	-0.04	-0.13	0.00	0.33
6	0.37	-0.43	+0.48	-0.44	-0.81	0.70	0.55
7	0.75	-0.18	-1.07	-0.17	+0.23	-2.87	0.24
8	0.82	-0.25	-0.62	-0.39	+0.49	-1.92	0.55
9	0.53	-0.60	-0.41	+0.49	-0.44	-0.14	0.56
10	0.18	-0.29	+0.59	-0.49	-0.24	1.58	0.73
11	0.63	-0.62	-1.12	+0.44	-0.90	-0.77	0.49
12	0.83	+0.34	-0.99	-0.24	-0.17	-2.88	0.37
13	0.85	+0.75	-0.97	-0.89	-0.16	-1.40	1.13
14	0.69	+0.49	-0.07	-0.45	-0.44	-1.03	0.55
15	0.60	-0.27	-0.61	+0.64	-0.41	-0.40	0.84
16	0.79	+0.28	-0.23	-0.49	-0.91	-2.95	0.29
17	0.43	-1.07	+0.42	-0.64	-0.60	0.45	0.47
18	0.78	-0.79	-0-34	+0.86	-0.47	-0.90	1.67
19	0.65	+0.76	-0.55	-0.62	-0.99	-0.50	1.17
20	0.26	-0.41	-0.23	-0.07	+0.34	1.86	0.36
21	0.44	-0.08	-0.08	+0.45	-0.52	0.33	0.51
22	0.37	+0.38	-0.64	-0.44	-0.16	0.89	0.41
23	0.71	-0.41	-0.30	-0.86	+0.81	-0.66	1.40
24	0.69	-0.19	-0.60	-0.34	+0.50	-1.01	0.56
25	0.59	+0.49	-0.86	-0.20	-0.45	-0.45	0.56

2.1.2. Test Parameters

The score formula for the test results is:

$$x_i = \sum_{g=1}^{25} U_g \qquad \left(U_g = \begin{cases} 0 \text{ for the wrong answer} \\ 1 \text{ for the correct answer} \end{cases}\right)$$

The mean ($\bar{x}$) of the values of x_i (i = 1, ... 68) is 14.72, and the standard deviation (s) is 4.21.

The test characteristic curve was built on the basis of the item characteristic curve. This curve is shown in Figure 5. As we can see, it is almost like a normal ogive and has a very good parameter of steepness. Therefore it can be concluded that the test is good; it has a good level of difficulty and a good discriminating power at the large interval of the latent attribute θ. The information function of the test (Figure 6) shows the same. The test is most

imformative and respectively effective for $\theta = 0$, where the value of $I(\theta)$ is 8.25 and its graph is similar to the normal distribution.

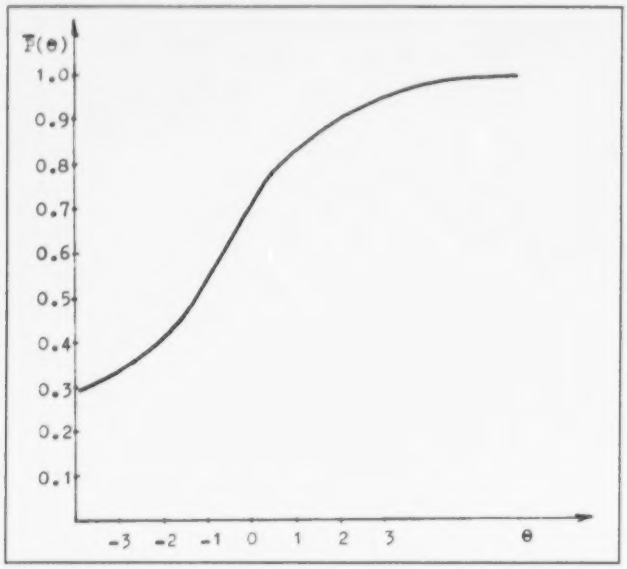


Figure 5. Test Characteristic Curve.

If we want to convert the raw score x_i into standard score, we can use Figure 5. In this case, we have to read the ordinate values as values of the proportion-correct score $z_i = x_i / 25$. For example, if $x_i = 15$, then $z_i = 15/25 = 0.6$, and from the graph in Figure 5 we can find that for $P(\theta)$ 0.60, $\theta = -0.70$.

The test reliability, according to Kuder-Richardson formula 20, is 0.74, and according to Kuder-Richardson formula 21 is 0.69. The difference between these two values is due to varying difficulty of the items. As the length of the test is relatively small, the value of the reliability coefficient is high enough. The standard measuring error is 2.15 and this too is a good result proving the test quality.

We also made some investigations of the criterion validity of the test. For these purposes we used two criteria: (a) the examinees' school marks in English, and (b) the test results from the test for verbal tropes understanding of the examinees.

The correlation coefficient between the test results and the marks is 0.39, and between the results from Test 1 and Test 2 is 0.52. These values are considerably greater than zero beyond the 0.001 level, but they are not high enough.

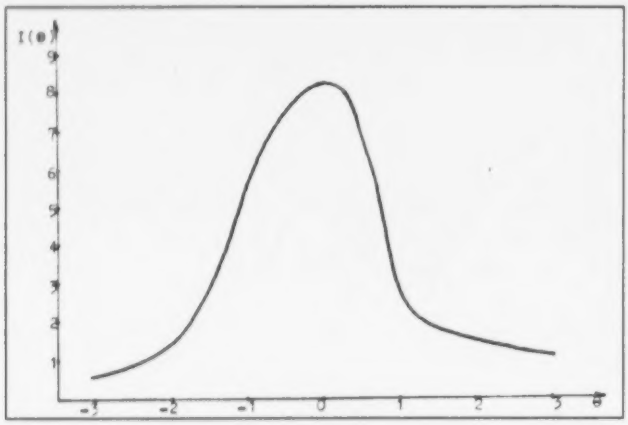


Figure 6. Test Information Function.

There may be various reasons for this but we think that the main reason is that the choice of criteria is not very good. We think that our test design correlates with the language abilities and general intelligence. That is why, future investigations on test validity will be made along these lines.

As a whole, the results we obtained on the basis of the tryout data give us grounds to think that the test has the required quality. For this reasons, we shall go on with the work on its validation, and respectively, its standardisation.

BIBLIOGRAPHY

Alderson, J.Ch. 1988. Testing English for Specific Purposes: How Specific Can We Get? in *Testing English for University Study*. Modern English Publications in association with the British Council.

Bachman, L. 1990. *Fundamental Considerations in Language Testing*. Oxford University Press.

Barthes, R. 1977. The Rhetoric of the Image; in *Image-Music-Text*. Fontana.

Clark, J. 1987. Curriculum Renewal in School. *Foreign Language Learning*. Oxford University Press.

Crystal, D. 1988. *The Cambridge Encyclopedia of Language*. Cambridge University Press.

Gak, V. 1974. *O funklionalnom podhode k izuceniju gramaticeskih javlenij v inostrannjie jazyku v vysej skole*.

Gelb, J. 1963. *A Study of Writing*. Chicago University Press.

Ivanov, V. 1978. *Cet i necet*. Assimetrija mozga i znakovyh sistem, Sovetskoe radio.

Lord, F. and Novick, M. 1968. *Statistical Theories of Mental Test Scores*. Addison-Wesley Publication Company.

Lossev, A. 1982. *Znak, simvol, mif. Izd*. Moskovskogo universiteta.

Mcluhan, M. 1965. *The Gutenberg Galaxy*. Toronto.

Moles, A. 1981. *L'image-communication fonctionnelle*. Casterman.

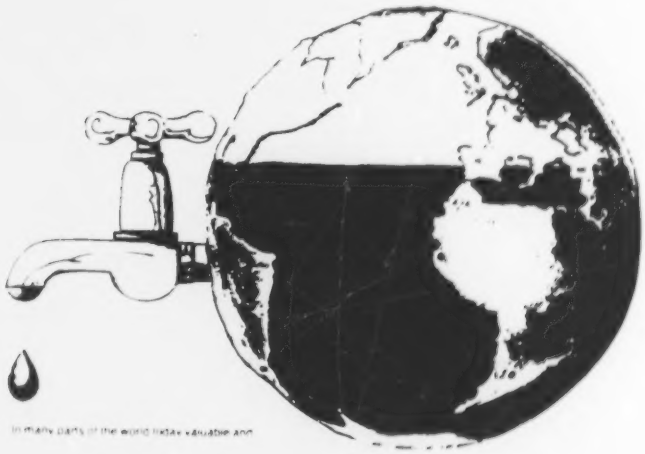
Morrier, H. 1961. *Dictionnaire de poetique et rhetorique*. PUF.

Taranenko, A. 1989. *Jazykovaja semantika v eje dinamiceskih aspectah*. Kiev, Naukova dumka.

Valette, R. 1971. Evaluation of Learning in a Second Language. In *Handboek on Formative and Summative Evaluation of Student Learning*. New York. McGraw-Hill Company.

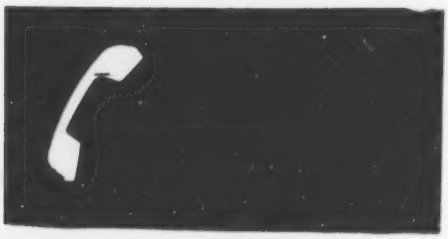
APPENDIX

1. Най-вярното твърдение е следното: 1



WE ALL KNOW IT'S TOO VALUABLE TO WASTE!
In many parts of the world today valuable and essential water is wasted and wastage caused by leaking and running taps is a major problem. VAMMALAN KONEPAJA INC. have developed a range of self-closing and automatically-closing taps that solve this problem.
А. Фирмата Вамалан е разработила нова система от резервоари, които биха помогнали за решаване на проблема с течовете.
Б. Фирмата Вамалан разработва някои проекти по отношение на екологични проблеми, свързани със засушаването.
В. Фирмата Вамалан се занимава с развойна дейност, свързана с нови типове кранчета за вода.
Г. Предметната дейност на фирмата Вамалан е свързана със създаване на уреди за премахване на дефекти по водопроводните системи.

2. Най-вярното твърдение е следното:



WE BECAME ONE OF THE LARGEST BUSINESS COMMUNICATIONS COMPANIES OVERNIGHT.
On July 7, 1988, three telecommunications industry leaders - Executone, ISOETEC and Vodavi - joined to form one exciting new company. A company that's one of the largest forces in the business communications market. Providing a full line of communications products, from sophisticated data switching equipment to individual phone stations. With 250 sales and service offices across the United States.
А. Нашата компания участва в разработката на комуникационни връзки за програмата Звездни войни.
Б. След обединението си нашата компания стана най-големата за производство на телефонни апарати за нуждите на бизнес комуникацията.
В. Нашата компания осигурява бизнескомуникациите чрез лансирането на телекомуникационни спътници.
Г. Нашата компания е една от най-големите в областта на комуникациите.

EST COPY AVAILABLE

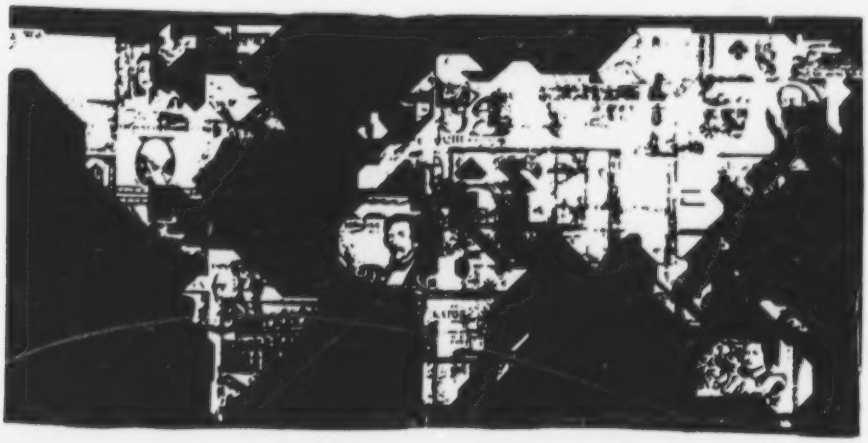
3. Най-вярното твърдение е следното: 2

SHARP-EYED KU RADAR SEEKS BRAINY COMPUTER FOR MARRIAGE.
Ideally a tracking radar should have a wide beam during acquisition and a narrow beam during tracking. Both of these phases are equally important for a successful engagement.
А. Нашите радари са със зоркостта на индиански вожд.
Б. Допълнителна екипировка е нужна на много добре виждащите радари на фирмата ни за докомплектуването им.
В. Радарът на нашата фирма е с принципно ново конструктурско решение.
Г. Радарите на фирмата ни търсят сътрудничество с истински воини.

4. Най-вярното твърдение е следното:

WE'RE TOUGH TO BEAT ON FIRM ORDERS
When the bell rings on your Firm Order, we come out fighting; and we won't quit until you have the books you ordered, on-time and at the best possible prices. We lead with worldwide English language coverage that includes all books published or distributed in North America; and that's backed by a proven 94% fill rate.
А. Ние ще направим всичко възможно, за да изпълним поръчките Ви за боксьорски инвентар на най-добрите възможни цени.
Б. Ние ще направим всичко възможно, за да изпълним поръчките Ви по отношение на спортна литература на най-добрите възможни цени.
В. Ние ще направим всичко възможно, за да изпълним поръчките Ви за всички видове литература на най-добрите възможни цени.
Г. Ние ще направим всичко възможно, за да изпълним поръчките Ви за спортни стоки по каталог на най-добрите възможни цени.

5. Най-вярното твърдение е следното: 3



GAIN WORLDWIDE INTEREST WITH A WISE INVESTMENT
When your objective is the mind of the international excutive, your financial statement should be in Newsweek International. The affluent. The educated. The English speaking, decision making business leader.

А. Ако иската да печелите на всички континенти в твърда валута, използвайте финансовите възможности на Нюзуик Интърнешънъл.
Б. Ако искате на внушите вашите идеи на финансовите босове, публикувайте в Нюзуик Интърнешънъл.
В. Ако искате да участвате в банковата игра по целия свят, вслушвайте се в препоръките на Нюзуик Интърнешънъл.
Г. Ако искате да сте сред първите по отношение оборота на паричната маса, използвайте Нюзуик Интърнешънъл.

6. Най-вярното твърдение е следното:

WEST LANDESBANK DERIVES ITS PROWESS IN INTERNATIONAL FINANCE FROM MULTIPLE SOURCES
Large-scale financing calls for a bank with all the credentials and expertise needed to ensure a smooth, competitive functioning of any major money raising operation.

А. Нашата банка развива многостранна дейност в най-различни клонове на икономиката.
Б. Нашата банка е смела в операциите си, защото има сведения от много източници.
В. Нашата банка извлича ценна информация за международното банково дело.
Г. Нашата банка финансира дейности, свързани с производството на енергия и топлина.

7. Най-вярното твърдение е следното:

JOIN THE SWARM
The C-212 M belong to a rare species. Renowned the world over for their capacity for work, they have another exceptional characteristic: versatility. The ability to handle the widest range of missions.

А. Нашият самолет е създаден за борба с различни видове насекоми.
Б. Нашият самолет изпълнява важни функции по отношение пръскането с инсектициди.
В. Нашият самолет няма подобен на себе си по изключителността на някои от своите характеристики.
Г. Нашият самолет е създаден да изпълнява много функции наведнъж.

8. Най-вярното твърдение е следното:

ARE YOU IN THE RACE FOR THE MOST VALUABLE DEVELOPMENT IN THE NORTH?
1300 acres of prime development land in the heart of a booming city. It sounds like a developer's dream. And this really is. Leeds is shaping up for the next century. Already, it is the best serviced motorway city in the UK and Britain's third most important centre for financial services.

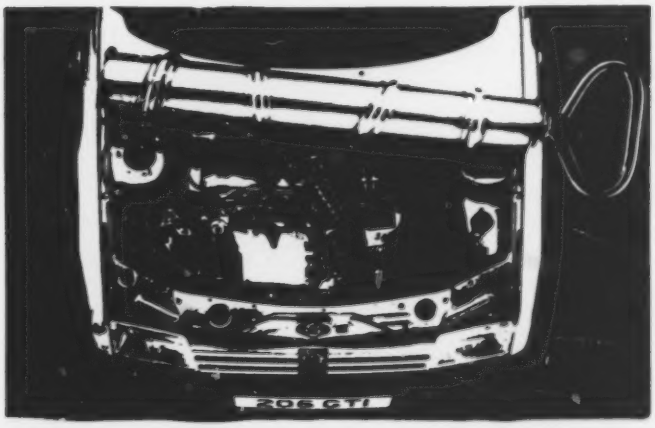
А. Състезавате ли се в различни дисциплини по отношение на обществения живот в Северния регион?
Б. Състезавате ли се за удоволствие и успех в някои дисциплини в Северния регион?
В. Състезавате ли се за печалба в някои видове спортове в Северния регион?
Г. Състезавате ли се за най-лакомото парче за инвестиране в Северния регион?

9. Най-вярното твърдение е следното: 5

UNLOCK THE DOOR TO EUROPE'S HIGH INTEREST RATES: CS EUROPE BOND
Yields on a number of traditional investment currencies are still low, despite the recent increases in interest rates. However, bond investments in many European countries now offer returns that are clearly superior the yields obtainable in hard currencies like DM, Dutch guilders, Swiss francs and yens. Private investors are therefore increasingly looking out for instruments that give them access to these attractive capital markets and involve a limited exchange and credit risk.

А. Частните инвеститори чакат премахването на бариерите пред финансовите операции.
Б. Частните капиталовложители търсят начини за излизането на пазарите.
В. Частните инвеститори търсят средства да влязат на най-привлекателните пазари.
Г. Частните капиталовложители търсят ключа за влагане на средства в някои специфични пазари.

10. Най-вярното твърдение е следното:



TO SEE HOW MUCH POWER SHE PACKS, TURN THE KEY AND OPEN HER UP
Peel back the slick, sculpted skin of the 205 GTi 1.9. Crammed beneath the bonnet, you'll find a heart designed with steely intent. A tightly woven construction of metal and silicon that delivers 130 BHP at 6,000 rpm and 121 lbs.ft.torque at 4,750 rpm.

А. Моделът Пежо 205, един от най-мощните, е в същото време и един от най-лесните за манипулиране по отношение на ред операции.
Б. Моделът Пежо 205 е с мощен мотор.
В. Моделът Пежо 205 е конструиран по нов начин по отношение отварянето на капака.
Г. Моделът Пежо 205 е решен по нови дизайнерски проекти.

11. Най-вярното твърдение е следното:

65

UNLIKE OUR HF HOPPING COMPETITOR, WE LOOK BEFORE WE LEAP.
Finally someone's come up with the technology to make tactical HF hopping really work. We have - Thomson-CSF. We call it Skyhopper and it's the first tactical HF concept that successfully gets around the problem of how do you hop frequencies when most of them are already occupied by someone else.

А. Нашата фирма е провела биологични изследвания и ги е приложила в производството на някои видове артикули.

Б. Нашата фирма е изучила някои от характеристиките на определени животински видове в търсене на екологично съобразени решения в някои производства.

В. Нашата фирма е разработила свързочна техника, която осигурява сигурна връзка, тъй като са намерени оригинални решения.

Г. Поведението на някои видове от животинската фауна може да се използва при създаването на продукти, използвани в същите екстремални биологични условия.

12. Най-вярното твърдение е следното:

TEN MILLION HIGH SCHOOL KIDS CAME FACE TO FACE WITH UGLY PROBLEMS
The poaching of the rhinoceros. It's just one environmental problem high school and junior high school kids are reading, thinking, and talking about in the pages of the six magazines that make up the The Scholastic Teen Network. Each day, school kids read about preserving the ecology of their palent, of their cities, and of their neighborhoods.

А. Десет милиона ученика четат за нерешените проблеми по отношение опазването на околната среда.

Б. Десет милиона ученика се сблъскват лице в лице с големи неприятности по отношение на училищния живот.

В. Десет милиона ученика се информират за проблеми, които са от ключово значение за тях.

Г. Десет милиона ученика преодоляват чрез списанието ни неприятностите, които са част от техния живот.

13. Най-вярното твърдение е следното:

HIGH VOLUME COPYING IS OUR FAVOURITE GROUND
The new Minolta copiers decisively improve on your handicap. Take the EP 8600. It quickly completes even the biggest copying tasks. Just press the button and the excitement starts: 60 sharp, crisp copies every minute.

А. Едно от основните преимущества на нашите машини е възможността за голям обем копия.
Б. Важно качество на машините ни е точното попадение по отношение качествата на копията.
В. Фаворитният спорт, спонсориран от нашата фирма, произвеждаща копирни машини, е голфът.
Г. Нашата фирма е сред фаворитите в състезанията за печелене на потребителите.

14. Най-вярното твърдение е следното:

FLY AT FULL STRETCH!
Now jet your way, cosy and comfortable, in one of our stretched first class seats - the sleeperettes of Kuwait Airways on Boeing 747s will get you fresh to your destination.

А. Кувейтските аеролинии предлагат като нова услуга - кресло, в което може да се леки изтегнат.
Б. Кувейтските аеролинии са едни от малкото в света, които предлагат толкова добро обслужване.
В. Кувейтските аеролинии правят всичко възможно пътниците да се чувстват напълно удобно в самолетите Боинг 747 ес.
Г. Кувейтските аеролинии откриват нова супер класа в самолетите Боинг 747 ес.

BEST COPY AVAILABLE

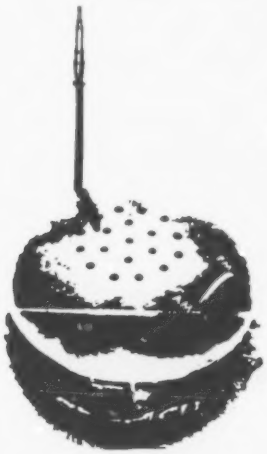
15. Най-вярното твърдение е следното:

AN OPPORTUNITY TO SHARE IN THE POTENTIAL OF EUROPE'S FINANCIAL SERVICES INDUSTRY: CS EURO FINANCE FUND
The creation of a single European market is adding a new dimention to the activities of Europe's financial services industry. The free movement of capital and the harmonization of the regulatory environment open up exciting new business opportunities.
А. Инвеститори влагат много средства в развитие на някои специфични видове индустрии.
Б. Финансови групировки на капиталовложители развиват дейности, които ще ви помогнат да се интегрирате в Европейската оптическа промишленост.
В. Капиталовложителите търсят подходящи средства, които ще им помогнат в сътрудничеството по отношение на финансови услуги.
Г. Компании от Европейската общност сътрудничат все по-тясно като банкови партньори.

16. Най-вярното твърдение е следното:

THE "4" ATTACHED TO THE NEW OVID-4 MEANS THAT AS OF NOW FOUR PROGRAMS CAN BE SIMULTANEOUSLY TRANSMITTED THROUGH ONE SINGLE GLASS FIBRE.
Those thoroughly familiar with the optical transmission of video signals are aware of the fact that the range of OVID equipment has already made television history.
А. Технологията позволява предаването на 4 програми едновременно.
Б. Технологията позволява предаването на 4 филма едновременно.
В. Технологията позволява заснемането и транслирането на четири фотографии едновременно.
Г. Технологията позволява големи възможности за манипулирането на 4 образа едновременно.

17. Най-вярното твърдение е следното: 9



MASS APPEAL
There are over half a billion radios in the United States. That means radio reaches everyone. Quickly and efficiently. Arbitron knows the mass appeal of radio. We measure the listening audience of stations all across America. And with all those radios, there's no doubt that America is listening.

А. На основата на проучването на потребителското мнение предлагаме радиореклами за нов тип много апетитен хамбургер.

Б. За милиони потребители има много привле..ателни неща и радиото е едно от тях.

В. Нашата фирма притежава национална верига от радиофицирани заведения "фаст фууд", които са привлекателни за милиони потребители.

Г. Милиони американци слушат предаванията на нашата агенция, предлагаща специфични типове услуги.

18. Най-вярното твърдение е следното:

4 SIZES FIT ALL. EASY COPYING, ENLARGING AND REDUCING
The U-Bix 330RE's unique 4-way cassette paper feeding system allows you to do some size copying, 2-step reducing and 1-step enlarging more efficiently because you don't have to change cassettes; they're already on line.

А. Нашите машини отговарят на изискванията по отношение на четирите основни критерия за оценяване на техниката на този клас.

Б. Нашите машини са най-сигурните по отношение осигуряването на различни варианти в някои видове бизнес.

В. Нашите машини осигуряват различна големина при репродуцирането на различни видове документи и други книжа.

Г. Нашите машини са едни от най-добре изпълнените и надежно работещите при проектирането

BEST COPY AVAILABLE

19. Най-вярното твърдение е следното:

NO ONE SAID IT WAS A PIECE OF CAKE
If your company uses telemarketing as a marketing tool, you understand why it works and maybe why it doesn't. In this business, hard work is not enough. Our profesaional recipe with quality, creative experience, and integrity delivers improved results. Telemarketing is tough - it's not a piece of cake.

А. Някои дейности от телемаркетинга са по-сложни отколкото изглеждат на пръв поглед.
Б. Само най-добре служещите си с телемаркетинга компании могат да се надяват на парче от тортата в своя бизнес.
В. Всяка фирма от някои браншове на промишлеността би могла да използва известна част от телемаркетинга.
Г. Бизнесът в сладкарската промишленост среща трудности по отношение на телемаркетинга.

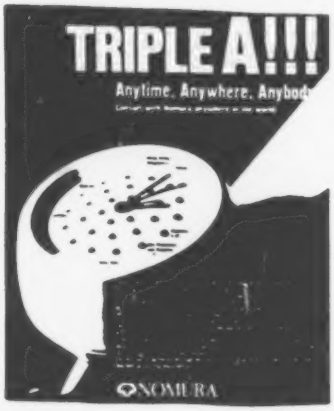
20. Най-вярното твърдение е следното:

CHAMPIONS OF POWER
The two newest champions in the Mitsubishi line up of diesel engines are the compact and lightweight, water-cooled Z Series and direct fuel-injected versions of the popular K Series engines. These multi-cylinder, fuel-efficient champions of power are the products of an innovative design concept that substantially reduces both engine size and weight and significantly increases the variety of usage and applications.

А. Нашите двигатели са най-пробивните на световните пазари в своята класа.
Б. Нашите двигатели са създадени от най-ударните конструкторски бюра на фирмата.
В. Нашите двигатели са най-защитените против повреди в своя клас.
Г. Нашите двигатели са най-силните в своя клас.

BEST COPY AVAILABLE

21. Най-вярното твърдение е следното:



TRIPLE A!!! ANYTIME, ANYWHERE, ANYBODY. CONSULT WITH NOMURA ANYWHERE IN THE WORLD. New York 10:00 A.M., London 3:00 P.M., Bahrain 6:00 P.M., Hong Kong and Singapore 11:00 P.M., and Tokyo 12:00 midnight. Somewhere, at this moment, there is important activity in the money market. In this age of international finance when trends in any one money marke of the world often have strong and immediate effect on all of the others, no investor can afford to limit his view to his own hometown.

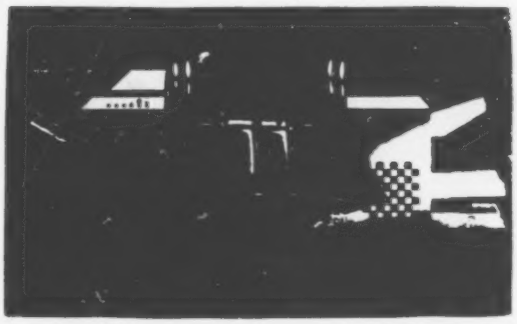
А. Нашата фирма осигурява бърза връзка в различни краища на света, независимо от местното време.

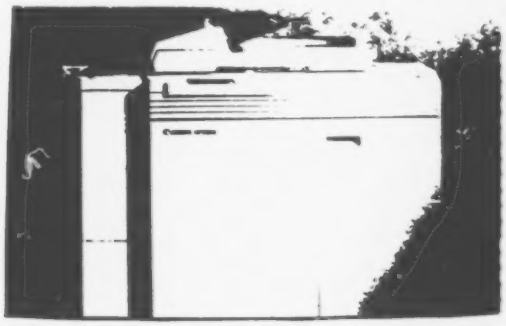
Б. Нашата фирма е с точността на швейцарски часовник, независимо в кой голям световен гра се намират нашите офиси.

В. Нашата фирма е специализирала в обслужване, свързано със съвети, в големите световни центрове по всяко време, когато това ви е необходимо.

Г. Нашата фирма е една от най-добрите в телекомуникационния бизнес - като производство и услуги навсякъде по света и непрекъснато.

22. Най-вярното твърдение е следното:





CATASTROPHE? OR CHAMPION?
Choosing a copier can be a risky business. At first sight, some macho machines look like they can stay the course. But test drive them in tough office conditions and you'll soon have a catastrophe. In contrast, the new Canon NP-6650 is a true champion. It can take the rough work with the smooth. At a superfast, 50 copies per minute (A4 size). Which is a record-breaking speed for such a compact copier.

А. Някои копирни машини са с твърде привлекателен вид, но нашите са истинските потребители.

Б. Някои копирни машини, рекламиращи с помощта на спортни автомобили, никога няма да станат шампиони.

В. Някои конструкторски решения по отношения на копирните машини са в тясна връзка с качествата на състезателните автомобили.

Г. Нашите копирни машини се тестират за големи скорости при вадането на копията с прецизността на тестиране на състезателните автомобили.

23. Най-вярното твърдение е следното: 1 2

TO JAPAN, NON-STOP. NOW! IPPON IT IS. THE FIRST AND ONLY NON-STOP FLIGHT FROM WESTERN EUROPE TO JAPAN.
To mark our 60th anniversary and in honour of His Majesty Emperor Hirohito's birthday, we are inaugurating the first non-stop air service between Western Europe and Japan on 29.4.1983. The flight over the North Pole will take 13 hours, leaving Helsinki on Fridays and Tokyo on Sundays.

А. Финландските аеролинии са победители в конкурентната борба между световните авиокомпании за директния маршрут Хелзинки - Токио.

Б. Финландските аеролинии покриват най-добрите японски стандарти по отношение на маршрута Хелзинки - Токио.

В. Финландските аеролинии, обслужващи маршрута Хелзинки - Токио, са спонсори на националния отбор по карате.

Г. Финландските аеролинии летят по маршрута директно Хелзинки - Токио.

24. Най-вярното твърдение е следното:

JAPAN AIR LINES HAS NEW POSITION FOR SUPER EXECUTIVE WITH BETTER HEARING, BROADER SHOULDERS, WIDER EXPERIENCE, AND OFFERING GENEROUS BENEFITS
From April, Japan Air Lines introduce a completely unique class of travel on all our B747s. Our new Super Executive "J" Class makes any executive really feel like a super executive. Because our executive "J" Class seats are wider. So there's more shoulder room. There's more sitting room. There's more padding for a softer life (and deeper sleep). There's even a foot-rest and more space for luggage.

А. Японските аеролинии въвеждат традиционни японски стандарти в самолетната класа Супер Игзекютив.

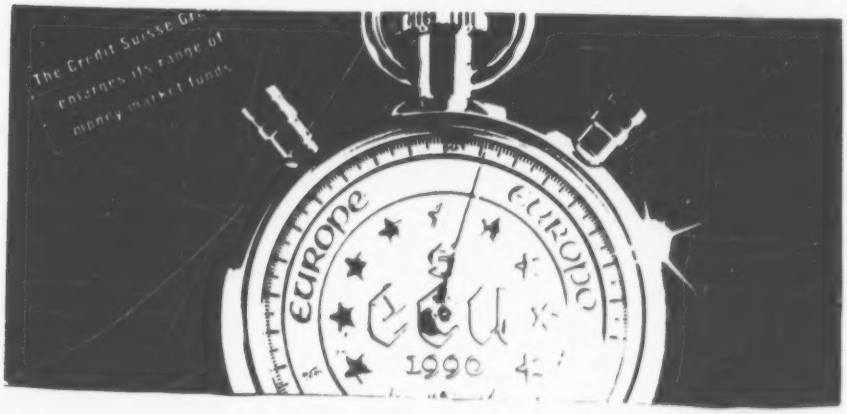
Б. Японските аеролинии поставят на първо място обслужването на пътниците в самолетната класа Супер Игзекютив.

В. Японските аеролинии осигуряват много удобно пътуване в самолетната класа Супер Игзекютив.

Г. Японските аеролинии през последните години осигуряват ново равнище на обслужване в самолетната класа Супер Игзекютив.

25. Най-вярното твърдение е следното:

1 3



NOW IS THE TIME TO INVEST IN CS MONEY MARKET FUND ECU
Inverted yield curves throughout the world are a strong argument for money market investments. ECU money market investments offer some particularly attractive features. By international standards, they carry a high yield, while the basket of currencies making up the ECU holds out the promise of a largely stable exchange rate.

А. Моментът за влагане в някои фондове е дошъл, тъй като фондовете предлагат много изгодни условия.

Б. Новата европейска валута дава големи възможности за инвестиране в часовникарската и измервателната промишленост.

В. Нашите банки са готови да инвестират във фонда "Екю" за успешното стартиране на новата европейска валута.

Г. Няма никакво време за отлагане на инвестирането на пари в специалните фондове, тъй като конкурентната борба е започнала.

RATING THE QUALITY OF STUDENT WRITING - FINDINGS FROM THE IEA STUDY OF ACHIEVEMENT IN WRITTEN COMPOSITION

Rainer H. Lehmann
University of Hamburg

The article reports findings on the reliability and validity of the writing assessments based on the IEA Study of Achievement in Written Composition. Major findings based on analyses of variance include the discovery of at least two different rating styles, an analytic and a holistic, and the difficulty of making generalizations about general writing ability on the basis of a limited number of writing assignments.

More often than not, research on the assessment of student writing has concentrated on the issue of inter-rater agreement (cf. the overview given in Breland, 1983). Given the importance attributed to writing assessment in many educational systems, the problem of minimizing rater influence carries, indeed, major practical significance. It should be noted, however, that the difficulties involved cannot be reduced to the problems of inconsistencies between the scores awarded by two or more independent raters. Rather, the practical value of writing assessments is also heavily dependent on the degree to which student compositions are representative of, or generalizable to, stable traits or abilities. This leads, of course, to the question of what is actually being measured in such assessments. As will be seen, the answers to all of these questions are intertwined in a rather complex manner.

1. DATA BASE

The data referred to here originate from the Study of Achievement in Written Composition, conducted by the International Association for the Evaluation of Educational Achievement (IEA; see Gorman, Purves, & Degenhart, 1988). This study is one of very few where students have been asked to write on several topics from the major "domains of school writing" (Vähäpassi, 1982), thus incorporating two to four texts per student to be analyzed:

-pragmatic texts, usually in a letter format;
-narrative, argumentative, and reflective texts;

-a "letter of advice", indicating students' perceptions of valued school writing.

The study is also unique in that juries of raters from 12 countries have worked on the basis of equivalent scoring schemes which involved a combination of overall impression marking and the use of analytical sub-scores under various rating aspects. Three of these analytical rating aspects are considered in the analyses to be reported here:

- quality and scope of content;
- structure and organization;
- style and appropriateness of tone.

In some countries, there are also data on up to three different age/grade populations, but the present considerations will be limited to the intermediate group, i.e. the final grade of compulsory schooling. The number of students and compositions rated in each country varied, depending on the study design chosen by each National Center. The exact numbers of cases will be given in the respective tables.

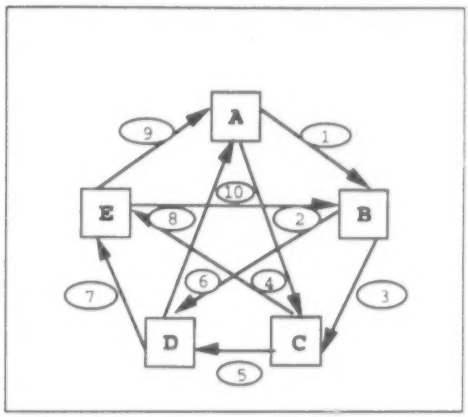
2. METHODOLOGY

When analyzing the quality of essay ratings, it is necessary to make inferences as to the following sources of variation

-method factors;
-between-rater factors;
-between-assignment factors;
-between-student factors;
-interaction effects between these four (cf. Wesdorp, Bauer, & Purves, 1982).

As stated above, prior research had concentrated largely on the second source of variation only. If all of these components are to be investigated simultaneously, the analysis of variance components appears to be an appropriate technique (cf. earlier suggestions by Gosling, 1966).

In order to visualize the structure of the data to be analyzed, two graphical representations may be helpful. Figure 1 gives an example of how one can distribute student essays among pairs of raters so that rater effects can be assessed best.



Rater	1st round package no.	2nd round package no.
A	1 2	8 9
B	3 4	10 1
C	5 6	2 3
D	7 8	4 5
E	9 10	6 7

Figure 1. Scoring scheme of the Hamburg Study "star design".

Such an arrangement makes sure that all possible combinations of raters occur in the data, should it be desirable to locate critical points of disagreement. At the same time, individual rater leniency effects (which would be critical in a large-scale survey) are structurally minimized. Figure 2 represents the formal structure of the data array thus obtained.

	First rating				Second rating			
Task Student	1A,B,C,D	5,6,7	9	0	1A,B,C,D	5,6,7	9	0
1	4	2	3	3	4	3	3	3
2	2	2	1	3	3	3	1	3
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
n	5	3	2	4	5	3	2	4

Figure 2. Analytic design and file structure of ratings.

It has been shown that parametric statistics render essentially the same reliability estimates as those produced by non-parametric methods which may justify to set aside fundamental objections against their use (Lehmann, 1990). A parametric approach does, however, have the advantage of allowing additional insights. More specifically, the techniques based on the analysis of variance (ANOVA) render a powerful set of tools, and within the domain of parametric statistics they prove to be generally superior to correlational methods. On this

basis, then, it is possible to derive from the classical definition of measurement theory,

$$\text{Reliability} = \frac{\text{true variance}}{\text{observed variance}}$$

precise specifications for a potentially unlimited number of situations. In particular, the problem of inter-rater agreement, which correlational statistics can handle only for pairs of raters, is then treated as a very simple special case.Those who are interested in the derivation of the respective formulae and in computational detail may wish to consult the excellent introduction provided in Thorndike (1982, 1988).

The general reasoning is as follows: Dividing the true variation, i. e., the observed total variation minus that which is considered as error variance, by the observed total variation renders the reliability of sum or average scores obtained from multiple ratings. The following equation holds:

$$\text{Reliability of multiple ratings} = \frac{MS_{\text{Essays}} - MS_{\text{Error}}}{MS_{\text{Essays}}}$$

Similarly, it is possible to estimate the reliability of a single component:

$$\text{Reliability of a single rating} = \frac{MS_{\text{Essays}} - MS_{\text{Error}}}{MS_{\text{Essays}} + (k - 1)MS_{\text{Error}}}$$

with k = number of ratings.

The latter measure is also known as "intra-class correlation", and it has been proposed occasionally as an appropriate indicator for the reliability of essay marking (Zeiher, Zeiher, & Krüger, 1979).

While this technique may be used to test the quality of just two ratings for any given essay, the major advantage of the ANOVA approach lies in the fact that it can be generalized to situations with more than just two factors involved. As was argued above, a systematic investigation of the quality of writing assessment requires more than just an analysis of rater agreement on a single rating aspect. The following will, therefore, be based on the technique of partitioning the observed variance into its components which — even if they be considered "error variance" — have a great deal to say about the character of the ratings produced by the IEA Study of Achievement in Written Composition.

3. THE RELIABILITY OF THE ANALYTICAL RATINGS

The recommendation to use multiple marking in the assessment of student writing is often justified with the argument that the individual ratings operate just like items on a test the more individual marks, the higher the reliability, because sums or averages tend to polish out at least some of the erratic components of the single mark. Not considering for a moment the problem of validity, the same argument can be used to defend the analytical scoring scheme applied in the study. If the raters can express analytically the particular strengths and weaknesses of an essay, the reliability of sums or averages calculated from the single scores will increase. It may also be, however, that the analytical scores for content, organization, and style contain true diagnostic information, which is likely to be the case when several raters deviate independently on a particular rating aspect in the same direction from the general tendency. In this situation, it would be clearly inappropriate to capitalize on the polishing effect of aggregated marks, and the total information should be used instead.

Which interpretation of the relative merits of analytical scoring fits the present data better can be investigated empirically. In order to do this, the ANOVA has to be specified as a three-dimensional model with essays as cases and Rating (first vs. second) and Rating Aspect (overall impression, content, organization, style) as facets. Again, the array of data contains a single code in each cell (zero within-cell variance), but now seven variance components are defined:

- Variation between *Essays*, computed from the averages of two ratings with four scores each (main effect E);

- variation between the *rating Aspects*, as resulting from averaging across the two rounds of scoring and the total *N* of essays (main effect A);

- variation between the *Ratings*, derived from the means of four scores and *N* essays (one for the first, one for the second round of scoring; main effect R);

- interaction between Essay and Rating Aspect (E x A), which can be interpreted, if high, as a sign of true diagnostic scoring;

- interaction between Essay and Rating (E x R), which is a measure of differential assessment between the two rounds of scoring, but across all four rating aspects;

- interaction between Rating Aspect and Rating (A x R), indicating aspect-specific differences in leniency between the two rounds of scoring;

- residual variation (E x A x R).

Table 1 presents each of these components (sigma^2) in percent of the total variance. These terms can be used in various ways to produce reliability estimates, depending on the assumptions one is prepared to make with respect to the role of the rating aspects and the raters (cf. the numerous variants presented in Thorndike, 1988, pp. 335-340).

TABLE 1. Achieved reliabilities for overall-impression scores (global marks) and analytical scores by country (all compositions with multiple ratings, irrespective of Task; constituent variance components in percent): Population B.

Country	*N* of composit	total variance	Overall impress.	single aspect	4 aspects random	4 aspects fixed
Chile	3087	.85	.82	.68	.81	.81
England	4317	1.23	.76	.68	.77	.77
Finland	3600	1.33	.80	.70	.80	.80
Hamburg	5286	.76	.89	.78	.90	.90
Hungary[a]	5689	1.19	.93	.83	.93	.93
Italy	5331	1.15	.66	.59	.68	.68
Netherlands[a]	3472	1.00	.77	.62	.79	.79
Sweden[b]	4101	1.18	.83	.76	.84	.84
(Sweden[c])	487	1.02	.82	.78	.85	.85
USA	1248	.78	.67	.59	.70	.70
Wales	1044	1.28	.61	.58	.66	.67
Median			.80	.68	.81	.81

Country	Essay	Aspect	Rating	Essay x Aspect	Essay x Rating	Aspect x Rating	Essay x Aspect x Rating
Chile	56	1	-	11	15	-	16
England	54	1	-	2	26	-	16
Finland	59	1	-	5	21	-	15
Hamb.	70	-	-	11	7	-	13
Hung[a].	71	1	-	10	6	-	12
Italy	43	-	-	3	32	-	22
Netherl[a]	42	1	1	10	17	-	29
Sweden[b]	63	-	-	3	19	-	16
(Swed[c])	56	-	-	3	25	-	17
USA	45	-	-	5	29	-	21
Wales	42	2	2	3	34	-	18
Median[e]	56	1	-	5	21	-	16

[a] compositions with triple ratings only
[b] all compositions with double ratings (Task 6 only, if third rating is missing)
[c] compositions with triple ratings only (Task 6)

An inspection of the table shows that only in Chile, Hamburg, Hungary, and the Netherlands the Essay-by-Aspect interaction term has reached or exceeded 10 percent of the total variance, while it is as low as 2 percent in some other countries, with an overall median of 5 percent or below. For most data sets, indications are that the analytical scoring has contributed little beyond providing an option to express special strengths and weaknesses, but with limited specific agreement among the raters.

This corresponds well with another observation. Rating Aspect has rarely produced a main effect associated with more than 1 percent of the total variation; indeed, it was to be expected that the raters would tend intuitively to aim at identical scale mid-points for all aspects (as was at least implicitly suggested by the given International Benchmark Set). Therefore, it is practically of no importance whether Rating Aspect is treated as a fixed or random factor. Both estimates (top) render generally higher values than the respective reliabilities of two independent "overall impression" scores alone, but since the average reliabilities for a single rating aspect (interpolated in the random model) are lower than the latter, the 'polishing effect' again appears to be the dominant advantage of the analytical scoring method used in the study.

A modified conclusion emerges from an investigation of those variance components associated with Rating. There is hardly any main effect for this facet, because the scoring designs have rather effectively eliminated this source of error, but the Essay-by-Rating interaction (*general* rater disagreement) and the residual term (*differential* rater disagreement) indicate substantial influence of rater disagreement on the achieved reliabilities. It is remarkable that *general* disagreement is in most countries the larger of these two components. Where this is not so —Chile, Hamburg, Hungary and The Netherlands — there is also a relatively large Essay-by-Aspect interaction (standing, as will be recalled, for rater *agreement* in diagnostic differentiation), and most of these data sets are associated with rather high reliabilities, even without considering the Essay-by-Aspect interaction as a source of true variation. So, it seems likely that at least here, the analytical categories have served to reduce effectively the general disagreement. The distinction between a more diagnostic, truly analytical rating behavior, which consciously tries to avoid within-essay halo effects, and a more holistic one, profiting from the 'polishing effect', might be referred to as one of *rating styles*. It will have to be borne in mind when looking at the problems of validity and generalizability of essay ratings.

Incidentally, it would have been easy to justify an estimation mode which combines the Essay-by-Rating component with the residual term, i.e., general and differential rater disagreement are then considered as a single source of error variation. This method would have produced slightly higher reliability estimates for the average score. It was preferred here to work within the more conservative — and more specific — approach. Likewise, the inclusion of the Essay-by-Aspect interaction component in the true variation would have resulted in higher achieved reliability estimates, and again the more

conservative option was chosen. So, the values given in Table 1 may be considered lower boundaries for the reliability of the ratings and their respective averages. Although future work with these data will have to take into account the considerable between-country differences, it may be informative to compare across-country medians of the achieved inter-rater reliabilities (Table 1) with estimates given by Breland in his summary of the state of the art (1983, p.10; cf. Table 2).

TABLE 2. Achieved reliabilities for aggregated scores in comparison with expected values reported by Coffman and Breland, by number of ratings.

N of ratings / Source	1	2	3
Coffman	.38	.56	.65
Breland	.64	.70	.78
IEA	.62	.76	.82

Source: Breland, 1983, Tables 5 and 6, and own calculations

Typically, the country contributions to the IEA study compare favorably with these standards.

4. THE VALIDITY OF THE ANALYTICAL RATINGS — SOME INDIRECT EVIDENCE

It has to be stated at the very beginning of this section The IEA files do not contain any variable against which to validate externally the ratings. So, it is not possible to study questions of the predictive or concurrent validities of the ratings, save by analyzing the relationships between ratings referring to different tasks. As will be seen, this approach produces at least some indirect evidence as to what has been measured. Table 3 presents the inter-task correlations.

There are considerable national differences between these correlations. Interestingly, higher reliabilities (as listed in Table 1) are not necessarily accompanied by higher inter-task correlations as one would normally expect. Even more intriguing, perhaps, there appears to be a rather close connection between the *rating styles of* national juries (as first defined in the context of discussing the results of Table 1) and the inter-task correlations. The more a national jury has emphasized the different rating aspects (as indicated by the Essay-by-Aspect interaction term), the lower are, at least tendentially, the inter-task correlations. This tendency is not as implausible as it may seem at first

Table 3. Inter-task correlations for overall-impression scores (averages from two ratings), by country: Population B.

Country: Task	Task 2	Task 3	Task 4	Narrat Task 5	Persua-sive Task 6	Reflect Task 7	Task 8	Task 9
Chile: Pragmatic (1)				.10	.42	.25		.30
Expository (9)				.40	.37	.44		
England: Pragmatic (1)			.44	.47	.50	.46		.52
Descriptive (4)				.40	.58	.54		.55
Expository (9)				.53	.61	.57		
Finland: Pragmatic (1)	.22		.41	.41	.43	.42		.46
Summary (2)				.34	.24	.28		.35
Descriptive (4)				.41	.35	.56		.57
Expository (9)				.50	.48	.64		
Hamburg: Pragmatic(1)				.18	.30	.29		.30
Expository (9)				.24	.25	.30		
Hungary[a]: Pragmatic(1)		.53	.41	.50	.43	.46		.58
Paraphrase (3)				.42	.55	.51		.61
Descriptive (4)				.36	.31	.43		.45
Expository (9)				.62	.58	.62		
Italy: Pragmatic (1)	.29		.25	.34	.24	.28	.44	.32
Summary (2)					.34			.41
Descriptive (4)				.46				.40
Reflective (7)							.51	.49
Expository (9)					.43	.49	.31	
Netherlands[a]:								
Pragmatic (1)				.35	.40	.45		.38
Expository (9)				.49	.55	.55		
New Zealand[b]:								
Pragmatic (1)			.36	.28	.28	.31		.36
Descriptive (4)				.48	.44	.47		
Expository (9)				.36	.39	.43		
Sweden[c]: Pragmatic (1)				.58	.59	.46		.52
Expository (9)				.59	.63	.59		
USA[b]: Pragmatic (1))				.23	.22	.32		.26
Expository (9)				.35	.40	.41		
Wales: Pragmatic (1)	.52		.61	.64	.70	.57		.58
Summary (2)				.56	.69	.57		.63
Descriptive (4)				.62	.60	.71		.77
Expository (9)				.67	.64	.65		

[a] averages from three ratings
[b] single ratings included
[c] third rating included, if available (Task 6)

sight; after all, the inter-task correlations were computed on the basis of the *global* "overall impression" scores, and, therefore, a more holistic approach to the rating of student essays is not unlikely to correspond better to this particular mode of investigating the meaning of the ratings across tasks. Nevertheless, there may well be more involved. It is possible that the difference between an analytically differentiated approach and a more holistic rating style implies a fundamental difference in the meaning of the judgments. While the former will have to stress *text characteristics*, the latter may — or, perhaps, is even likely to — capitalize more on stable *student* traits. This question will be taken up again.

It is also clear, however, that the nature of the tasks involved has also had some influence on the inter-task correlations. While it does not appear that these differences can be easily related to the theoretical grid of "domains of school writing" (Vähäpassi, 1982) originally developed to provide a framework for the study, some observations do stand out.

Task 9 - the "letter of advice to a friend" spelling out what needs to be done to obtain good marks for one´s writing - is special, indeed. It was administered with relatively little time allowed for completion (20 minutes), but it is the nature of this task which makes it so interesting. To articulate in writing one´s own perceptions of the relevant criteria of good writing requires a *reflective mode*, and it will be fairly safe to expect raters to award high marks on this task only if an essay provides significant evidence of this. (Content analytic techniques applied to the Hamburg data have shown that this is, indeed, the case.) Thus, it is quite remarkable to find that this particular task is generally associated with the highest inter-task correlations and that the highest coefficient within the array for a country is usually the one connecting Task 9 with the reflective Task 7! It seems that these findings confirm the notion of a school-specific *writing culture* The fact that the ratings in general are highest for those students who know best how to express their teachers' criteria as to good writing is probably an indication for the effectiveness of mutual expectations between students, teachers - and raters.

As far as the magnitude of the coefficients is concerned, the inter-task correlations are more or less in line with what could be expected or even slightly higher. In his state-of-the-art report of 1983, Breland has compiled the results from eight studies on the predictive and concurrent validities of essay scores (pp. 13 f.). His tables would provide medians of .23 for concurrent correlations with verbal intelligence tests, etc., and of .30 for predictive correlations, e.g., with course marks (excluding the secondary analysis of Werts et al., 1980), which leads to rather modest expectations to begin with, even if it is also clear from this material that correlations will tend to be higher to the extent that the validation criterion resembles the used indicators of writing achievement.

Very sporadically, the concurrent validity of analytical rating aspects has been investigated differentially (cf. Breland, p. 17, Table 11). In such studies, higher

values were found regularly for language-related characteristics (style and grammar, for instance) than for measures of content quality and structure. This finding is confirmed by the IEA ratings. If rating aspects are interpreted as 'traits' and writing assignments as 'methods' to measure them, it is possible to compute "convergent validity coefficients" (Campbell & Fiske, 1959; Marsh, 1988). As can be seen from Table 4, these are generally higher for Style than they are for Content and Organization.

TABLE 4. Convergent validity coefficients for averages from two or more ratings, by country and rating aspect: Population B.

Rating aspect / Country	*N* of task combi-nations	Overall impress.	Content	Organiz.	Style and Tone	$\bar{X}$
Chile	7	.3238	.2233	.2812	.2740	.2756
England	12	.5158	.4171	.4631	.4988	.4737
Finland	17	.4096	.3667	.4032	.4283	.4020
Hamburg	7	.2666	.2132	.2308	.3144	.2563
Hungary[a]	17	.4924	.3723	.4320	.4714	.4420
Italy	16	.3745	.3161	.3510	.3534	.3488
Netherlands[a]	7	.4544	.2426	.3071	.3739	.3445
New Zealand[b]	11	.3768	.3187	.3720	.3916	.3648
Sweden[c]	7	.5655	.4659	.4713	.5154	.5045
USA[b]	7	.3133	.2602	.2729	.2738	.2800
Wales	17	.6306	.4659	.5117	.6469	.5638
Median		.4096	.3187	.3720	.3916	

[a] triple ratings
[b] including single ratings
[c] including third rating where available (Task 6)

That leads to the question of the construct validity of the analytical scoring scheme as a whole. The labels for the rating aspects in the International Scoring Guide, common across all tasks, and parallel phrasings in the respective circumscriptions seem to suggest that these aspects are invariant against text genres and tasks. This is an assumption, however, which warrants testing. It was not possible to do this for all data sets, but principal component analyses could be performed on the ratings from Hamburg and Sweden in order to clarify the factorial structure of 18 variables, six each from three tasks (Task 1 vs. Tasks 5, 6, 7 combined vs. Task 9). The rating aspects included were Overall Impression, Content, Organization, Style, Grammar, and Handwriting/Neatness. A slightly modified version on the Hamburg analysis is shown in Table 5.

TABLE 5. Pattern matrix from a factor analysis on 40 variables, referring to four different assignments (pairwise deletion of missing observations, oblique rotation, loadings < .30 not shown; *N*=1340)

Variable	Factor	1	2	3	4	5	6
Pragmatic Task	Overall			.92			
	Content			.86			
	Organization			.89			
	Style			.82			
	Grammar			.36			.40
	Spelling						.60
	Handwriting/Neatness					.75	
Narrat./ Argumentat./ Reflective Task	Overall		.96				
	Content		.95				
	-Quality		.90				
	-Scope		.93				
	Organization		.92				
	-Overall structure		.90				
	-Control of detail		.91				
	Style		.87				
	-Tone		.91				
	-Choice of words, phrases etc.		.81				
	Grammar						.55
	Spelling						.67
	Handwriting/Neatness					.79	
Letter of Advice	Overall				.90		
	Content				.87		
	Organization				.87		
	Style				.79		
	Grammar						.52
	Spelling						.67
	Handwriting/Neatness					.75	
Narrat./ Argumentat./ Reflective Task	Overall	.94					
	Content	.91					
	-Quality	.86					
	-Scope	.91					
	Organization	.91					
	-Overall structure	.89					
	-Control of detail	.91					
	Style	.87					
	-Tone	.88					
	-Choice of words, phrases etc.	.82					
	Grammar	.40					.41
	Spelling						.62
	Handwriting/Neatness					.82	

The respective results were quite unambiguous. Apart from Handwriting and Grammar, the factors are clearly task factors. In fact, one can go even one step further. In an optional exercise, the essays referring to text analyses (Hamburg only) were rated for a second time, evaluating the three sub-tasks (paraphrasing; analysis of rhetorical means; elaboration of judgement) separately. The factor pattern matrix (Table 6) is equally clear, save the interesting observation that the ability to paraphrase appears to be a rather general trait.

TABLE 6. Pattern matrix from a factor analysis on 25 variables, referring to three different sub-tasks of Task 0 (pairwise deletion of missing observations, oblique rotation, loadings < .30 not shown; *N*=1340)

Variable	Factor	1	2	3	4
Overall			.44	(.26)	
Paraphrase	Overall		.83		
	Content		.78		
	Organization		.88		
	Style		.72		
Rhetorical Analysis	Overall			.91	
	Content			.96	
	-Quality			.90	
	-Scope			.92	
	Organization			.73	
	Style		.36	.56	
Personal Judgment	Overall				.89
	Content				.94
	Organization				.86
	Style		.33		.62
Complete Task (original ratings)	Overall	.88			
	Content	.86			
	-Quality	.77			
	-Scope	.88			
	Organization	.89			
	-Overall structure	.84			
	-Control of detail	.91			
	Style	.96			
	-Tone	.87			
	-Choice of words, phrases etc.	.87	(.27)		

Incidentally, in the comparison of factor solutions between Hamburg and Sweden (not shown here), the different weights/eigenvalues of the first principal component demonstrated again the existence of different rating styles. Apparently, the Swedish jury has been guided more strongly by characteristics of language (also visible from the fact that there is no independent grammar

factor here, but high loadings of Grammar on the task factors instead), while the Hamburg jury differentiated more between Content/Organization and Style and made "quality and scope of content" decisive for their "overall impression" scores. It was also implied by the results that stylistic characteristics and the ability to write grammatically correct are relatively near to each other, even if this relation was almost entirely masked by halo effects in Sweden. Taking it all together, however, these findings confirm basically the distinction of Quellmalz, Capell, & Chou (1982) between a task-specific "Coherence" factor (content and organization) and an independent "Mechanics" factor. Conversely, this replication of earlier results suggests a high degree of validity attributable to the IEA ratings.

5. THE GENERALIZABILITY OF THE RATINGS

The IEA study comprised quite a number of different tasks, and it is rare in the field of writing assessment to find a corpus with more than a single text from each student, such as here. The tasks given do not stand for themselves. They were developed to cover as wide a range of school writing as seemed possible under the anticipated time limits and under the constraints of an international project with very diverse participating school systems. They were also meant to cover as wide a range of performance characteristics as possible. So the question is almost unavoidable. To what extent can the essay ratings be generalized across tasks into a single estimate of student writing achievement?

Technically, this question can be treated as a special topic of the problem of reliability. It can be investigated adequately with the aid of variance components analysis, if there are at least two sets of ratings per student. It was deemed appropriate here, however, to discuss first the question of the validity of the analytical ratings, because it follows directly from these analyses (as was already suggested by the findings from Section 3) that average scores computed from the rating aspects Overall Impression, Content, Organization, and Style are adequate measures of the quality of an essay as perceived by the raters.

So, the technique of characterizing the specifics of each national rating file by its constituent variance components can now be resumed, but with a change in the statistical design. The cases (= unit of analysis) are now students, and the facets are Task (strictly speaking "assignment", because rotated tasks had to be combined in order to have a complete block design) and Rating. In accordance with the overall design of the study, both Task and Rating (first vs. second round of scoring) are treated as random factors. It is to be noted that only the averages across the four main rating aspects are entered into the computations.

Now, the following variance components are defined:

- Variation of *Student achievement*, computed from the averages across two, three, or four tasks and two independent ratings (main effect S);

- Variation associated with differences between the mean ratings for the different *tasks*/assignments (main effect T);

- Variation between first and second *Rating* (main effect R);

- Student-by-Task interaction, interpretable as an indicator for fluctuations in the performance level across tasks to the extent that they were perceived by *all judges* (S x T);

- Student-by-Rating interaction which indicates systematic changes in scoring standards for certain (groups of) students between scoring rounds (S x R);

- Task-by-Rating interaction, a measure of differential changes of standards at the task level between scoring rounds (T x R);

- Residual variation which encompasses all 'erratic' rater disagreement (S x T x R).

Table 7 presents the respective data; the generalizability coefficients (top right column) were computed in accordance with the stated specifications, and the variance components (bottom) are again the sigma^2 terms converted to percentages of the total variance which is also given.

As was stated above, the problem of the generalizability of essay ratings can be understood as a special topic in the study of reliabilities; it represents a necessary, however rarely pursued question: What is the degree of precision with which we can make inferences as to a single, general measure of writing achievement, if we assume the tasks/assignments to be a random sample from the universe of all possible tasks considered relevant? An inspection of the figures given in the respective columns of Table 7 must have sobering effects. If "general writing achievement" (or "general writing ability", for that matter) is a meaningful concept — and many programmatic texts, e.g., preambles of mother tongue curricula, seem to start out from that assumption — and if the measurement of *general* writing achievement was an important aim of the study, then this has been achieved only moderately well.

TABLE 7. Achieved generalizabilities for averages across four rating aspects on the basis of multiple writing assignments ("essays") and multiple ratings by country (constituent variance components in percent), Population B.

Country	*N* of assignm.	*N* of ratings	*N* of students	Total variance	Generaliz-ability
Chile	2	2	733	.69	.38
England	3	2	1097	1.01	.76
Finland	3	2	671	1.11	.70
Hamburg	3	2	1169	.65	.55
Hungary	3	3	942	1.12	.68
Italy	3	2	645	.94	.62
Netherlands	3	3	972	.70	.65
Sweden[a]	3	2	1283	1.01	.72
Median					.68
England	4	2	473	1.04	.77
Finland	4	2	340	1.10	.75
Hamburg[b]	4	2	1088	.62	.62
Hungary	4	3	373	1.00	.72
Italy	4	2	203	.89	.70
Median					.72
Country (*N* of essays)	Stud.	Essay	Student x Essay	Student x Rating	Student x Essay x Rating
Chile (2)	23	-	62	4	10
England (3)	42	-	22	-	37
Finland (3)	38	4	27	1	31
Hamburg (3)	26	5	57	-	12
Hungary[c] (3)	34	10	28	1	13
Italy (3)	28	-	23	1	48
Netherlands (3)	29	4	30	1	33
Sweden[a] (3)	42	5	27	-	27
Median	34	4	27	1	31
England (4)	38	-	23	1	38
Finland (4)	37	3	27	1	32
Hamburg (4)	29	3	56	-	11
Hungary (4)	36	11	41	-	12
Italy (4)	29	-	18	1	52
Median	36	3	27	1	32

[a] all compositions with double rating (Task 6 only, if third rating is missing)
[b] including nationally optional task
[c] also 4 percent for rater component and 11 percent for essay x rater component

When looking for causes, it is helpful to analyze the variance components contributing true and error variance. It is obvious that inter-rater disagreement (the residual term) is only one — and in some countries, a rather minor — source of error variance. Clearly, the fluctuations of student performance between tasks (S x T interaction) have a major influence — and in some countries they constitute the largest variance component of all. Since the generalizability coefficient takes into account all these components, it is easily seen that a low level of rater disagreement *and* a low level of within-student achievement variation are required to obtain better results. More texts would also have helped, of course, and it may be worth remembering that the nature of the assignments does seem to be related to their power of allowing valid predictions of achievement in different assignments (cf. the discussion of Table 3).

The rating files as such do not allow to identify external reasons for the observed student fluctuation of achievement over tasks. It is well possible that context factors are involved where larger student-task interactions are observed, e.g., class- or school-specific opportunities to learn. So this must be left for further investigations. It is, however, possible and even likely that the *rating styles* which have emerged from the previous analyses are also to be considered here. When comparing the Essay-by-Aspect interaction terms from Table 1 with the Student-by-Task interaction terms from Table 7, there is an almost perfect correlation of rank orders, and, of course, the between-country differences in inter-task correlations have illustrated basically the same phenomenon. This, then, suggests the following tentative explanation. The clearer the distinction between content and language, i.e., variable task requirements and general characteristics of 'good language', the lower the overall emphasis on student traits which are relatively stable across the different tasks.

If this interpretation is correct, it has far-reaching implications for the study of writing achievement. It follows directly that all correlations with a measure of writing achievement as the criterion will depend on the specific rating style of the jury and, further, that the explained variances in multiple regression analyses and causal models will be immediately affected. The IEA data do seem to confirm this hypothesis (see Purves, in print). In the light of these results, comparisons of writing assessments across different juries appear to be even more problematic than had been anticipated. At the same time, the very fact that considerable variation in rating styles existed between national juries has added to the existing knowledge of the constituents of composition ratings.

6. CONCLUSIONS

The reported analyses of the international rating files from the IEA study allow to draw the following conclusions:

1. The internationally standardized scoring procedures produced ratings from independently and "blindly", i.e., uninformed as to any student characteristic, operating scorers, which — for most countries and populations are reliable in accordance with the state of the art. Most rating files fit the model of analytical scoring as a way to provide repeated, equivalent measures which increase the overall reliability when averages of the existing rating aspects are used.

2. Some indirect evidence could be produced as to the validity of the analytical scoring scheme. Whereas writing achievement must be primarily defined with respect to specific tasks completed, there are traces of a distinction between content-related and language-related text characteristics.

3. For most countries and populations, it does not seem justified to aggregate ratings across tasks into a single measure of 'general writing achievement'. Unexplained variations of writing achievement within-students, between-tasks lead in most cases to rather low generalizability estimates.

4. The national juries differ clearly in what has been suggested here to be called "rating style". Some juries have distinguished more clearly than others between the analytical rating aspects. This has had demonstrable effects on the inter-task correlations. It is hypothesized that a more pronounced analytic rating style with its emphasis on content qualities of the text as opposed to language ability of the student will be associated with lower effect sizes, as far as student-related predictors are concerned.

REFERENCES

Applebee, A. N., Langer, J. A., & Mullis, J. V. S. (1986) *The Writing Report Card.* Writing Achievement in American Schools. (Report No. 15-W-02). Princeton, NJ (Educational Testing Service.

Breland, H. M. (1983). *The Direct Assessment of Writing SkillA Measurement Review.* (College Board Reaport No. 83-6; ETS RR No. 83-32). Princeton, NJ (Educational Testing Service).

Campbell, D. T. & Fiske, D. W. (1959) Convergent and discriminant validation by multitrait-multimethod matrix. *Psych. Bull.*, 56, pp.81-105.

Coffman, W. E. (1966) On the validity of essay tests of achievement. *Journ. of Educ. Measurement*, 3, pp.151-156.

Diederich, P. B., French, J. W., & Carlton, , S. T. (1961) *Factors in Judgments of Writing Ability.* (Research Bulletin RB 61-15). Princeton, NJ (Educational Testing Service).

Diederich, P. B. (1966) How to measure growth in writing ability. *English Journal*, 55, pp.435-449.

Gorman, T. P., Purves, A. C., & Degenhart (Eds.) 1988. The *The IEA Study of Written Composition I The International Writing Tasks and Scoring Scales.* (International Studies in Educational Achievement, vol. 5). Oxford (Pergamon Press).

Gosling, G. W. H. 1966. *Marking English Composition.* Hawthorn, Victoria (Australian Council of Educational Research).

Lehmann, R. H. 1990. Aufsatzbeurteilung - Forschungsstand und empirische Daten. In Ingenkamp, K. & Jäger, R. S. (Eds.)*Tests und Trends.* Jahrbuch der Pädagogischen Diagnostik, vol. 8. Weinheim (Beltz Verlag), pp. 64-94.

Marsh, H. W. 1988. Multitrait-multimethod analysis. InKeeves, J. P. (Ed.) *Educational Research, Methodology, and Measurement.* An International Handbook. Oxford (Pergamon Press), pp.570-580.

Purves, A. C. 1987. The organisation of raters in marking writing assessments. In Degenhart, R. E. (Ed.) *Assessment of Student Writing in an International Context.* (Publication Series B: Theory into Practice, Vol. 9). Jyväskylä (Institute for Educational Research), pp. 107-113.

Purves, A. C. (ed.) 1992. *The IEA Study of Written Composition II.* Oxford (Pergamon Press).

Quellmalz, E. S., Capell, F. J., & Chou, C. 1982. Effects of discourse and response mode on the measurement of writing competence. *Journ. of Educ. Meas.*, 19 ,(4), pp.241-258.

Thorndike, R. L. 1982. *Applied Psychometrics.* Boston, Mass. (Houghton Mifflin).

Thorndike, R. L. 1988. Reliability. In Keeves, J. P. (Ed.) *Educational Research, Methodology, and Measurement.* An International Handbook. Oxford (Pergamon Press), pp.330-343.

Törmäkangas, K. 1987.The validity, reliability, and generalizability of ratings of written composition. In Degenhart, R. E. (Ed.) *Assessment of Student Writing in an International Context.* (Publication Series B: Theory into Practice, Vol. 9). Jyväskylä (Institute for Educational Research), pp. 107-113.

Vähäpassi, A. 1982. On the specification of the domain of school writing. *Evaluation in Education*, 5, (3), pp.265-289.

Werts, C. E., Breland, H. M., Grandy, J., & Rock, D. R. 1980. Using longitudinal data to estimate reliability in the presence of correlated measurement errors. *Educ. and Psych. Measurement*, 40, pp.19-29.

Wesdorp, H., Bauer, B. A., & Purves, A. C. 1982. Toward a conceptualization of the scoring of written composition. In *Education in Evaluation*, vol. 5, pp. 299-315.

Zeiher, H., Zeiher, H. J.,& Krüger, H. 1979. *Textschreiben als produktives und kommunikatives Handeln.* Untersuchungen und Konzepte zum Deutschunterricht. Bd.I. Stuttgart (Klett).

A MODEL OF ESL READING COMPREHENSION DIFFICULTY

Kyle Perkins and Sheila R. Brutten
Department of Linguistics
Southern Illinois University

Variables which affect the probability that an English as a Second Language (ESL) student makes the correct choice on a standardized, group-administered reading comprehension test were examined in this study. Percent of content words per passage, the number of predicates per passage, passage predicate density, and cognitive demand were the variables which uniquely accounted for a significant amount of variance in the item difficulty for each item.

INTRODUCTION

The purpose of this study was to identify the variables which uniquely accounted for a significant amount of variance in the percent correct obtained by an ESL subject pool for each item in a standardized, group-administered reading comprehension test. The criterion variable will hereafter be referred to as item difficulty, or p value. The rationale for the study was motivated by the following considerations. First, identifying the variables for predicting performance on reading comprehension tests is a major focus in reading research. Second, there are many potential sources of difficulty in a reading comprehension test, and these sources of difficulty may derive from the way in which the prose passages and the reading comprehension questions are constructed (Scheuneman, Gerritz & Embretson, 1989). Third, test item writers do not usually control nor quantify the sources of difficulty, and apparently little is known about what factors account for the observed difficulty of multiple-choice reading comprehension test items (Embretson & Wetzel, 1987). Fourth, subject matter and testing experts cannot reliably estimate the difficulty of a test item (Bejar, 1983).

In this paper we will begin with a discussion of a current theoretical model of reading comprehension. The second section discusses the notion of difficulty in terms of reading comprehension. The third section presents the operationalization of the variables tested in the study. The next section describes the subjects and the procedures of the study. The fifth section presents the statistical results. The last section discusses the findings.

1. A THEORETICAL MODEL OF READING COMPREHENSION

Any model of reading comprehension must acknowledge at least three essential components: a text, a reader, and the reader's interpretation of the text. In addition, the type of text utilized in the reading process and the reader's formal and content knowledge of the text, and the reader's purpose(s) for reading must also be considered (Calfee & Drum, 1986).

There are many operations that go on during the reading process which must be accounted for by a model of reading comprehension. Some of these operations include word recognition, meaning decoding, segmenting sequences of words into grammatical constituent structures, deriving statements from combinations of meaning, "inferring connections among statements, holding in short-term memory earlier concepts while processing later discourse, inferring the writer's intentions, schematization of the gist of a passage, and a memory retrieved in answering questions about the passage" (Bower & Morrow, 1990, p.44).

It is generally agreed that comprehension involves translating the text into propositions, identifying the referents of the propositions, inferring causal relations and linking relationships, constructing a mental model of the text, and updating the model (Bower & Morrow, 1990).

Earlier models of the reading comprehension process favored a top-down orientation which emphasized knowledge-based, context-driven expectations and hypothesis testing. More recent models are empirically oriented, are concerned with modeling data, and can be referred to as connectionist architectures (Stanovich, 1991a, 1991b).

An example of a model which rejects expectancy-based, content-driven, top-down processing is Kintsch's (1988) construction-integrations model of reading comprehension. Kintsch's construction phase consists of: "(a) forming the concepts and propositions directly corresponding to the linguistic input; (b) elaborating each of these elements by selecting a small number of its most closely associated neighbors from the general knowledge net; (c) inferring certain additional propositions; and (d) assigning connection strengths to all parts of elements that have been created" (p.166). Construction is a bottom-up process wherein a text-based proposition network is derived and connected to knowledge structures. "In the integration phase, activation spreads through the network and stabilizes in a connectionist manner to determine a coherent interpretation. In the construction-integration model, text information contacts and shares activation with knowledge structures, but comprehension is not 'driven' by knowledge-based expectations in the traditional top-down fashion" (Stanovich, 1991a, p.78).

1.1. Difficulty in Reading Comprehension

The surface structure variables of a text, its question stems, and response alternatives have been shown to account for item difficulty in paragraph comprehension items (Drum, Calfee, & Cook, 1981). The variables included the number of content words and sentence length, among others. The importance of surface structure variables is related to the reader's limited processing capacity and the fact that a reader cannot focus attention on more than one thing at a time. A large number of sentences per passage and a large number of test questions per passage could possibly overload the reader's processing capacity and hence affect the likelihood of choosing the correct answer.

Readers generally perform better on passages with which they have familiarity both with the content and formal structure. Thus, passage content has been shown to have a strong association with differential item functioning (Scheuneman & Gerritz, 1990).

Propositional analysis has enjoyed wide-spread utilization as a means for representing text processing difficulty and for describing the semantic content of prose passages (Kintsch & van Dijk, 1978). The number of propositions in a sentence determines reading time; therefore, texts with relatively more propositions should be more difficult to process. One would then expect proposition density to be related to item difficulty (Embretson & Wetzel, 1987). In fact, research has shown that texts containing many different arguments, rather than fewer arguments, required longer reading times (Kintsch, Kozminsky, Streby, McKoon, & Keenan, 1975). The propositional content and density must be examined in both the passage and item stems, because both are sources of potential difficulty.

Reading comprehension test items may assess text representation abilities (the reader's process of converting text stimuli into a meaningful representation) and decision processing abilities (the reader's process of comparing the question stem and the choice alternatives to the text for accuracy) (Embretson & Wetzel, 1987; Embretson, Fultz, & Dayl, 1989). There are four decision processes which could pose difficulty for readers.

The reader must examine the text to search for the relevant propositions to answer the question. This process is known as text mapping, and it concerns the proportion of text over which the relevant propositions for answering a question are located. "If the relevant propositions are contained in a single short sentence, it is expected that text mapping would be easy. However, if the relevant propositions are spread through the text, or if the alternative concerns the `gist' of the paragraph, text mapping would be more difficult" (Embretson & Wetzel, 1987, p.178).

There are two stages in which readers evaluate the truth status of the alternatives (Pellegrino & Glaser, 1979; Whitely & Barnes, 1979). The first stage

is falsification. If the distractors are not wholly false, if the truth status of the distractors cannot be determined, and if readers can falsify few alternatives, an item will be difficult.

After falsification, readers compare the remaining alternatives to the text to ascertain whether the text confirmed the correct answer for a question. This process is known as confirmation. If the text offers few, if any, clues for confirmation, an item will be difficult.

Decision processing also involves propositional comparison during which readers compare propositions to falsify or confirm an alternative. If the alternatives contain several propositions, and each of the propositions must be compared to several propositions in the text, then an item will be difficult.

The cognitive demand which reading comprehension questions place upon readers can be yet another important source of difficulty. In terms of cognitive demand, reading comprehension items can be categorized into two categories; identification items which entail the identification, recognition, naming, discerning, locating, exemplifying, or illustration of a concrete piece of relevant information in the text (a lower-order cognitive process) and non-identification items which entail inferencing and generalization (a higher-order cognitive process). Research has shown that information may be processed at varying depths and that identification, lower order questions require less cognitive processing than questions which require more than direct memory, and further that it is possible for a reader to answer an identification question by matching the stem and responses with the text.

On the other hand, higher-level, non-identification questions such as generalization and inference questions, for example, require readers to evaluate, generalize, infer, and interpret relationships between the elements of a text which are not explicitly stated. Higher-order questions such as generalization and inference questions must be processed at a deeper cognitive level than at the direct memory level, and they require readers to attend to and to recall more of the passage than its surface orthographic or phonological features. In other words, to answer a verbatim factual question, readers need to retain the elements of the question in direct sensory memory long enough to match those elements with the surface orthographic or phonological features of the pertinent section of the text. On the other hand, to answer a higher-level question such as a generalization or an inference question, readers must retain the environmental sensory information in short-term memory long enough for it to be encoded and stored for later transfer to the long-term memory (Perkins & Brutten, in press).

1.2. Operationalization of the Variables Tested in the Study

There were four classes of variables examined in this study: (1) various counts of text surface structure (Drum, Calfee, & Cook, 1981), (2) propositional analysis of the passages and item stems (Scheuneman & Gerritz, 1990; Scheuneman, Gerritz, & Embretson, 1989), (3) decision processing (Embretson & Wetzel, 1987; Embretson, Fultz, & Dayl, 1989), and (4) cognitive demand (Scheuneman, Gerritz, & Embretson, 1989).

At this juncture, it is necessary to make the following points. First, the rather large list of predictors was chosen, because each has been thoroughly researched in first-language studies, but not in second-language studies; thus it was not possible to state research hypotheses, because there is no basis in second-language reading research on which to base predictions. The theoretical structure for the list of predictors is the success of the variables for predicting item difficulty in first language studies. Second, it is not possible to demonstrate that the reading comprehension questions utilized in this study test the processes outlined in the theoretical model of reading comprehension, because the elicitation instrument is a secure test, and the item descriptions are proprietary.

Text Structure. The variables describing the structure of the texts included the number of paragraphs per passage, the number of lines per passage, the number of test items per passage, the number of words per passage, the number of content words per passage, the number of sentences per passage, a passage/word sentence ratio, and the percent of content words per passage.

Under the category of text structure we also included a passage content variable: if a passage was humanities based, it was coded 1. A non-humanities passage was coded 0.

The number of test items per passage was placed under Text Structure, because we could find no other suitable generic category for it.

Propositional Analysis. We employed Scheuneman and Gerritz's (1990) approach to propositional analysis, namely identifying and counting the words indicating the presence of (a) predicates (all verbs, other than those only indicating tense; members of objects; negatives; and possessives), (b) arguments (uniquely appearing subjects and objects), and (c) modifiers (adjectives and adverbs) (p.114). The propositional analysis for the passages and item stems was separated, because predictive power is lost when information is combined (Scheuneman, Gerritz, & Embretson, 1989). The different types of propositions were separated, because research has shown that using a total propositional count considerably reduces item difficulty prediction (Embretson & Wetzel, 1987).

The following propositional counts were conducted for both the passages and the item stems separately: the number of arguments, the number of modifiers, the number of predicates, argument density (the number of arguments divided by the number of sentences), modifier density, predicate density, and combined density (the total number of propositions divided by the number of sentences).

Decision Processing. For this study, text mapping was operationally defined as the proportion of text over which the relevant propositions for answering a question are located. For the falsification process, the difficulty was calibrated as "1.0 minus the probability that the correct answer could be selected by random guessing among the remaining non-falsified set" (Embretson & Wetzel, 1987, p.179). For coding confirmation, if the text confirmed the correct answer for a question, a 1 was recorded. If the text did not confirm the correct answer, a 0 was recorded. Propositional comparisons were measured by "the natural logarithm of the number of propositional comparisons required to falsify (or confirm) an alternative" (Embretson & Wetzel, 1987, p.179).

Cognitive Demand. Scheuneman, Gerritz, and Embretson (1989) used five cognitive process categories in their analysis, and we modified those categories as follows for our study:

0= identify, recognize, name, discern, locate, match, exemplify, or illustrate a concrete piece of relevant information in the text which was given in an item stem.

1= non-identification
- support/weaken a claim, procedure, or outcome; substantiate, demonstrate, prove, confirm, verify a result; negate, critique, contradict, or disprove a claim, procedure or outcome
- infer, conclude, induce, deduce, diagnose, distinguish, differentiate, contrast
- generalize, plausibly universalize, find common ground, transfer, analogize, apply, carry over
- problem/solve, calculate, inquire, experiment, evaluate, appraise, weigh, compare

(Adapted from Scheuneman, Gerritz, & Embretson, 1989, pp. 13-14).

If a test item required the reader to conduct a cognitive operation on a concrete, verbatim piece of information in the text, that item was coded 0. All other cognitive requirements were coded 1. The decision was made to collapse the processes into a dichotomous variable because not all non-identification processes were manifested in the test which we used as an elicitation instrument, and frequently there was only a single item for a particular non-identification process, making any sort of statistical analysis meaningless.

1.3. Construct Validity

The previous sections on a theoretical model of reading comprehension and difficulty in reading comprehension have attempted to explicate the sources of text surface structure, propositional analysis, decision processing, and cognitive complexity that can be related to item difficulty in a reading comprehension test. The purpose of the study was to test the theoretical model of ESL reading difficulty by using statistical procedures and comparing the resulting statistical model of item difficulty to the hypothesized model in the test of reading comprehension utilized in the study.

2. METHOD

2.1. Participants

Seventy students enrolled in an intensive English language center participated as subjects for this study. The distribution of native languages in the subject pool was the following: Japanese, sixteen; Chinese, thirteen; Arabic, twelve; Korean, eleven; Spanish, nine; Thai, two; Turkish, two; Urdu, two; Hebrew, one; Indonesian, one; and Woolf, one.

2.2. Instrumentation

The elicitation instrument was 29 items from the *Test of English as a Foreign Language*, Form 3LTFG (Educational Testing Service, 1990). The TOEFL reading comprehension test consists of short reading passages each of which is followed by a set of items.

2.3. Data Coding

Each of the twenty-nine reading comprehension test items was coded according to four sets of variables: text structure, propositional analysis of passages and stems, decision processing, and cognitive demand. Each co-author coded the items independently and anonymously. Disagreements were adjudicated by a third party. Coding stability consistency was determined by calculating Pearson correlations for the continuous variables and percentage agreements for the categorical variables. The coefficients ranged from .85 to .93.

2.4. Dependent Variable

The dependent variable in the study was the item difficulty of each item, commonly referred to as the p-value, or the proportion of correct responses. The p-value for each item was transformed to the delta scale. Delta is a normal

deviate with a mean of 13 and a standard deviation of 4. The delta scale yields a statistic that is almost always positive ranging in practice from 5 to 24.

2.5. Statistical Analysis

The statistical analysis consisted of two phases. First, the variables were correlated with the item difficulty indices. The Pearson was used for the continuous variables, and the point-biserial correlation was used for the dichotomous variables. Only those variables which correlated with item difficulty at the .05 level were retained for further analysis. Second, only those variables which correlated significantly with item difficulty at the .05 level were submitted to a stepwise regression analysis. The purpose of the regression analysis was to determine how much variance each variable uniquely accounted for in item difficulty.

3. RESULTS

Table 1 presents the means and standard deviations of the variables. Under the heading, Test Data, we included the average number of correct responses, the average total TOEFL score, and the average classical test theory item difficulty to give readers an idea of the proficiency level of the subject pool. These data did not enter the analysis, and they are for information purposes only. The individual item difficulties were transformed into standardized units, averaged, and then the average was converted back to an item difficulty index. The average item difficulty is a little above, i.e. .07, the mid-point on the item difficulty continuum which ranges from 0.00 to 1.00.

Table 1. Means and Standard Deviations of the Variables.

Variable	Mean	S.D.
Test Data		
Correct responses (29 possible)	16.56	4.46
Total TOEFL scores	476.66	45.03
Item difficulty (p-value, non-delta)	0.57	0.21
Text Structure		
Passage content (1= humanities; 0= non-humanities)	0.45	0.51
Number of paragraphs per passage	1.60	0.89
Number of lines per passage	16.40	5.39
Number of test items per passage	6.00	1.41
Number of words per passage	232.80	70.88
Number of content words per passage	140.00	43.78
Number of sentences per passage	10.60	4.39
Passage word/sentence ratio	22.72	4.08
Percent of content words per passage	0.60	0.05

Propositional Analysis: Passage		
Number of arguments per passage	57.80	16.35
Number of modifiers per passage	46.40	23.74
Number of predicates per passage	23.80	5.07
Passage argument density	5.67	1.04
Passage modifier density	4.36	1.52
Passage predicate density	2.39	0.55
Passage combined density	12.41	2.43
Propositional Analysis: Stem		
Number of arguments per stem	3.07	1.09
Number of modifiers per stem	2.28	1.16
Number of predicates per stem	1.76	0.83
Stem argument density	1.08	0.32
Stem modifier density	0.70	0.39
Stem predicate density	0.66	0.36
Stem combined density	2.53	0.52
Decision Planning		
Text mapping	0.19	0.13
Falsification	0.72	0.16
Confirmations	0.59	0.50
Propositional comparisons	5.57	1.61
Cognitive Demand		
(0= identification, 1= non-identification)	0.62	0.49

Table 2 portrays the correlation coefficients for all the variables analyzed in the study. Five variables correlated significantly ($p<.05$) with item difficulty: cognitive demand; the number of test items per passage; percentage of content words per passage; passage predicate density; and the number of predicates per passage. These five variables were the only variables retained for the stepwise multiple regression analysis.

Table 2. Correlations with Item Difficulty.

Cognitive demand	.603	**
Number of test items per passage	.587	**
Percent of content words per passage	.569	**
Passage predicate density	.484	*
Number of predicates per passage	.381	*
Number of paragraphs per passage	.277	
Confirmations	-.275	
Falsifications	-.234	
Number of modifiers per passage	.229	
Number of content words per passage	.222	
Passage modifier density	.222	
Number of arguments per stem	-.209	
Propositional comparisons	.206	
Passage argument density	-.205	

Passage combined density	-.152
Stem predicate density	.133
Number of words per passage	.128
Stem argument density	-.123
Stem combined density	.110
Number of predicates per stem	.102
Stem modifier density	.099
Number of sentences per passage	.091
Number of lines per passage	.082
Passage word/sentence ratio	.066
Text mapping	.064
Passage content	-.051
Number of arguments per passage	.016
Number of modifiers	-.011

Critical tabled values:
**r= .487, p <.01, two-tailed test
*r= .381, p <.05, two-tailed test

Table 3 presents the results of the regression analysis. Cognitive demand entered the equation first and accounted for 35.84 percent of the variance in the item difficulty. The number of test items per passage entered the regression model second accounting for an additional 15.56 percent of variance. Passage predicate density and the number of predicates per passage entered the equation at steps three and four, and the two variables together accounted for an additional 4.86 percent of variance in item difficulty (2.14 + 2.72 = 4.86). At step five the program removed the number of test items per passage and replaced that variable with percent of content words per passage which accounted for an additional 15.91 percent of variance. At step six the model re-entered the number of test items per passage. The entire model accounted for 72.49 percent of variance in the reading comprehension test.

Table 3. Regression Analysis Results. R-square Improvement for Dependent Variable Item Difficulty.

Variable Entered	R-square	R-square Increase	F	Prob>F
Cognitive demand	0.3584	0.3584	15.08	0.0006
Number of test items per passage	0.5140	0.1556	8.33	0.0078
Passage predicate density	0.5354	0.0214	1.15	0.2930
Number of predicates per passage	0.5626	0.0272	1.49	0.2337
Number of test items per passage removed				
Percent of content words per passage entered	0.7217	0.1591	29.64	0.0001
Number of test items per passage	0.7249	0.0032	0.26	0.6153

Table 4 presents the results of the best five-variable model. This table shows which variables uniquely accounted for a significant amount of variance in the dependent variable. As Table 4 indicates, only four variables uniquely accounted for significant variance in the dependent variable, item difficulty: passage predicate density, the number of predicates per passage, percentage of content words per passage, and cognitive demand. In terms of sets of variables, the significant predictors of item difficulty came from text structure, propositional analysis: passage, and cognitive demand. No significant predictors of item difficulty were found in the propositional analysis: stem and decision planning sets of variables.

Table 4. Best Five-variable Model Showing Unique Variances Accounted for by the Predictor Variables.

R-square = 0.7249

	DF	Sum of Squares	Mean Square	F	Prob>F
Regression	5	325.2804	65.0560	12.12	0.0001
Error	23	123.4430	5.3670		
Total	28	448.7233			

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
Intercept	147.6689	36.6511	87.1249	16.23	0.0005
Passage predicate density	-6.1871	1.5203	88.8798	16.56	0.0005
Number of predicates per passage	-0.5629	0.1446	81.3188	15.15	0.0007
Percent of content words per passage	-144.6866	39.2863	72.7966	13.56	0.0012
Cognitive demand	-3.3516	1.1925	42.3938	7.90	0.0009
Number of test items per passage	- 0.3595	0.7058	1.3927	0.26	0.6153

4. DISCUSSION

Throughout the discussion, the fact that data were elicited from only seventy subjects from only one test form, the preliminary nature of the results, and the need for further refinement should be kept in mind. That is, we do not imply that the results from this study are generalized to all ESL/EFL populations using any reading comprehension test.

Predicate propositions and passage predicate density had a significant relation in the model. The explanation for this finding is that many predicates in the list of propositions appeared to function as superordinates, participated frequently in processing cycles, and linked the text for coherence. As the number of predicates and their density increased, so did the item difficulty. Dense terms of action and being were significantly related to item difficulty.

In the previous paragraph we mentioned that many of the predicates appeared to participate frequently in processing cycles. Processing cycles is a construct from the Kintsch and van Dijk (1978) model of text comprehension which is described as follows:

> The reader processes text in cycles, during which a network of coherent propositions is constructed. During a cycle, certain propositions are retained in a limited short-term buffer for connection with input on the next cycle. An input is accepted as coherent with previous text if a connection is found between new propositions and those retained in the buffer. The reader may encounter difficulty during the cyclical construction of a coherent text base if no connection of the input to previously processed propositions is encountered, causing memory search, rereading, or generation of inferences to bridge gaps in the text. (Embretson & Wetzel, 1987, pp. 177-178).

Percent of content words per passage also contributed to item difficulty: the higher percentage of content words in the passage, the more difficult the items associated with that passage. This variable relates directly to the amount of semantic information readers must process; therefore, higher semantic content per reading passage would be expected to make comprehension more difficult. Embretson and Wetzel (1987), in a discussion of percent content words, noted that "the introduction of new material or new word concepts should also place greater demands upon the reader's resourcefulness" (p. 189).

Cognitive demand uniquely accounted for a significant amount of variance in item difficulty: the identification items which entailed the identification, recognition, naming, discerning, locating, exemplifying, or illustration of a concrete piece of relevant information in the text which was given in an item stem were systematically associated with easy items. Those non-identification test items which entailed higher-order cognitive processes of inference, generalization, etc. were associated with the difficult items. There has been research conducted to explain why identification items are easier to answer than the non-identification items found in the test. Information may be processed at varying depths and that the identification-type questions employed in this research require less cognitive processing than questions which require more than direct sensory memory (Perkins & Brutten, in press). The cognitive demand variable bundles more basic elements, which may be the reason that it predicts so well. This is a subject for further research.

At the outset of the study we were convinced that argument propositions and their density would be significant predictors of item difficulty; they were not. The lack of a significant effect for arguments is probably due to the fact that we did not distinguish superordinate and subordinate arguments. Superordinate arguments relate to global meanings; subordinate arguments relate to local meanings. Superordinates contain information relevant to the entire text and link text for coherence; subordinates do not. Because the reading passages were relatively short, we did not operationalize the superordinate and the subordinate distinction of arguments. Kintsch and Keenan (1973); Kintsch, Kozminsky, Streby, McKoon, and Keenan (1975); and Miller and Kintsch (1980) have demonstrated that structural superordinates are recalled better than structural subordinates, possibly because superordinates are processed more frequently and contain information relevant to the entire text. The operationalization of the superordinate/subordinate distinction will be a concern for future research.

The decision process set of variables did not correlate significantly with item difficulty and therefore were not entered into the regression analysis. Because the TOEFL is a speed test (timed), not a power test (non-timed), our ESL subject pool probably did not have enough time (1) to map the alternatives to the text, (2) to find the relevant text for evaluating each alternative, and (3) to evaluate the truth status of the alternatives. These three events are thought to comprise the decision process, according to the Embretson and Wetzel (1987) paradigm.

The results indicate that for the subject pool which was previously described and the elicitation instrument employed in this study, difficult items have (1) texts that have a high proportion of content words, (2) many different predicates densely nested, and (3) cognitive demands that require higher-order processing.

In summary then, the purpose of this study was to test a theoretical model of ESL reading difficulty which derived from first-language reading research by using statistical procedures and comparing the resulting statistical model of item difficulty to the hypothesized model in a second-language test of reading comprehension. The intent was to shed new light on the nature of ability by explicating the text structure, propositional analysis, decision process, and cognitive demand that are involved in item solving. The results indicated relatively good prediction in that the model accounted for 72.49 percent of the test variance with percent of content words per passage, the number of predicates per passage, passage predicate density, and cognitive demand each uniquely accounting for significant variance.

REFERENCES

Bejar, I. 1983. Subject matter experts' assessment of item statistics. *Applied Psychological Measurement*, 7, 303-310.

Bower, G.H., & Morrow, D.G. 1990. Mental models in narrative comprehension. *Science*, 247, 44-48.

Calfee, R., & Drum, P.A. 1986. Research on teaching reading. In M.C. Wittrock (Ed.), *Handbook of research on teaching* (Third edition) (pp.804-849). New York: Macmillan Publishing Company.

Drum, P.A., Calfee, R.C., & Cook, L.K. 1981. The effects of surface structure variables on performance in reading comprehension tests. *Reading Research Quarterly*, 16, 486-514.

Educational Testing Service. 1990. *Test of English as a Foreign Language*, Form 3LTF6. Princeton, NJ: Author.

Embretson, S.E., & Wetzel, C.D. 1987. Component latent trait models for paragraph comprehension tests. *Applied Psychological Measurement, 11*, 175-193.

Embretson, S., Fultz, J., & Dayl, N. 1989. The influence of paragraph comprehension components on test validity. In R.F. Dillon & J.W. Pellegrino (Eds.), *Testing: Theoretical and applied perspectives* (pp. 36-65). New York: Praeger.

Kintsch, W., & Keenan, J. 1973. Reading rate and retention as a function of the number of propositions in the base structure of sentences. *Cognitive Psychology*, 5, 257-274.

Kintsch, W., Kozminsky, E., Streby, W.J., McKoon, G., & Keenan, J.M. 1975. Comprehension and recall of text as a function of content variables. *Journal of Verbal Learning and Verbal Behavior*, 14, 196-214.

Kintsch, W., & van Dijk, T.A. 1978. Toward a model of text comprehension and production. *Psychological Review*, 85, 363-394.

Kintsch, W. 1988. The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review*, 95, 163-182.

Pellegrino, J.W. & Glaser, R. 1979. Cognitive correlates and components in the analysis of individual differences. *Intelligence*, 3, 187-214.

Perkins, K., & Brutten, S.R. (in press). The effects of processing depth on ESL reading comprehension. *Journal of Research in Reading*.

Scheuneman, J., Gerritz, K., & Embretson, S. 1989, March. *Effects of prose complexity on achievement test item difficulty*. Paper presented at the meeting of the American Educational Research Association, San Francisco.

Scheuneman, J.D., & Gerritz, K. 1990. Using item differential item functioning procedures to explore sources of item difficulty and group performance characteristics. *Journal of Educational Measurement*, 27, 109-131.

Stanovich, K.E. 1991a. Cognitive science meets beginning reading. *Psychological Science*, 2, 70, 77-81.

Stanovich, K.E. 1991b. Word recognition: Changing perspectives. In R. Barr, M.L. Kamil, P. Mosenthal, & P.D. Pearson (Eds.), *Handbook of reading research* (Vol. 2) (pp. 418-452). New York: Longman.

QUESTION LEVEL BIAS IN CLOZE QUESTIONS - AN L1 TRANSFER EFFECT

Alastair Pollitt and Lynda Taylor
Research Centre for English and Applied Linguistics
University of Cambridge

The article reports the findings from item analyses carried out on the (productive) cloze test used in the Cambridge First Certificate in English examination. Most of the items were found to be biased towards one or more of the L1s studied; the reasons for this are discussed. The writers argue that the cloze test studied was essentially a test of productive language competence which probably produces undesirable bias effects when used with candidates coming from different L1 backgrounds.

INTRODUCTION

The study reported here arose as a complication in a larger project investigating the construct validity of a cloze test used in Cambridge English proficiency examinations. It raises interesting and potentially serious issues for the use of cloze in multi-L1 contexts.

1. CONSTRUCT VALIDATION

The traditional approach to the **construct validation** of a test has been to try to explain why some people get higher scores on it than others do. Indeed Cronbach (1984) actually defines the term in this way as:

Construct Validation: Explaining Test Scores and Their Relations

Traditional studies concentrated exclusively on explaining the variation of person scores; in terms of the usual test data matrix these are the row sums of the data. But the data matrix can equally well be summed vertically, giving column sums, or item scores. It is surprising how little attention has been paid to the variation of item scores as evidence about what a test measures. The test is, after all, made up of items, and the principal statistic or parameter used to describe items — all test theories agree on this — is a measure of item score, or easiness, or difficulty. The most direct way to answer the question of what a test

measures would seem to be to look at the questions, to try to see what people have to do to answer the questions, to see what is difficult in them.

Stenner et al. (1983) offer some other reasons for studying item score variance rather than person score variance to validate a test construct. In normal circumstances a small number of items are given to a relatively large number of people; this means that item scores will typically be known more accurately than person scores, and so will provide a firmer basis for theory development. Unlike people, items do not change with time; a theory can therefore be developed without necessarily collecting new data, since the investigator can devise and apply new measures of the same items at any time in the future, and study their effects in the original item/person data set.

Several researchers in the late 1970's studied the 'cognitive components' of test items, mostly from tests of visual, spatial or other simple cognitive abilities (see, for example, Wright & Stone (1979), Embretson (1983)). Stenner et al. (1983) used such an approach to explain the difficulty of items in the Peabody Picture Vocabulary Test. A large number of characteristics of the test items (53, including interactions and polynomials) were used as predictors in a multiple regression analysis of Rasch item difficulties. From this, a best set of three predictors was obtained, explaining 71% of the variance; it consisted of log word frequency, dispersion of word across contexts, and abstractness.

This work may be described as essentially empirical, since there was little or no theoretical justification for the choice of potential predictors. Bachman et al. (1991) used Bachman's (1990) CLA and TMF frameworks to generate the item and text characteristics to be used as independent variables. Some 30 variables were used in an unconstrained regression analysis to predict the difficulty of Cambridge proficiency items. The results were mixed, with virtually no prediction when the vocabulary and reading items in six FCE examinations were combined, but 56% success when reading items alone were predicted.

A different theoretical framework was used by Embretson & Wetzel (1987) to analyse multiple choice reading comprehension items. It was assumed that test takers comprehend the passages in terms of Kintsch & van Dijk's (1978) propositional analysis, and then choose their answer by comparing propositions in options and text until each option is confirmed or refuted. Predictors were derived from the densities of various types of propositions in the text, and from the number of propositional comparisons between question and text required to reach a definite answer. Approximately 37% of the question difficulty variance was explained in this way.

In all of these studies the text — and the question — seem to be treated as objects to be analysed exhaustively and logically by the test taker. As Brown & Yule (1983) say of Kintsch's analysis,

> *'..the solution proposed by Kintsch may be an acceptable heuristic in experimental psychology, but it can have only an extremely limited application in discourse analysis.'*

The problem is that there is no unambiguous procedure for propositionalising a given text, as Kintsch himself admitted (Kintsch, 1974:124); and plenty of evidence suggesting that each reader constructs, to some extent, their own interpretation, as a result of the nature of their particular schemata and of the particular inferences that they make.

It is not at all clear that test takers do analyse text in a propositional way, and it seems that a rather different **psycho-logic** may be a more appropriate heuristic.

In a 1983 study, Pollitt et al examined what makes examination questions difficult. In one part of the study, reported more fully in Pollitt & Hutchinson (1986), a psychological process model was constructed for the answering of open ended passage comprehension questions in both first and second languages. It was proposed that the test taker goes through the following steps:

1 **Reading** and understanding the question
2 **Searching** for the relevant part(s) of the text
3 a **Decoding** the relevant text words
  b **Comprehending** the relevant text words in relation to the question
4 **Composing** and writing an answer

Analysis of the errors made by 550 candidates allowed each of these steps to be elaborated by identifying as many as possible potential sources of difficulty in questions of this type. The result was a 'flow diagram' containing 44 question variables. Some clusters of these variables were combined into single variables, so that the final set consisted of 22 question variables arising in a strictly linear sequence. 62% of the question difficulty variance was predicted.

The advantage of modelling the process of answering questions with a strictly linear sequence of variables is that it allows a rigorous interpretation of the regression analysis, with an unambiguous test of the significance of each variable. The increase in R^2 as each variable is added measures

> ...its ability to add extra predictive power over and above that already provided by the variables that precede it. (Pollitt & Hutchinson, 1986; 54)

The work reported here is part of a project aiming to develop a similar linear model of the process of completing a cloze test.

2. PROCEDURE

2.1. Rasch Estimation

A 20 item cloze test from the December 1990 Cambridge First Certificate in English examination was chosen for the study, and a sample of 300 scripts selected. 30 scripts were selected, from several examination centres, in each of the Netherlands, Germany, Norway, Spain, Brazil, France, Turkey, Japan, Hong Kong and Bahrain/Oman. Each sample was screened to try to ensure that every candidate was a native speaker of the majority language in that country; any one whose name suggested the possibility that they were not was replaced. The ten groups may therefore be taken to represent ten different L1 groups. The test is given at the end of this paper.

There are three parallel parts to the analysis. First, Rasch analysis of the data established difficulty values as the dependent variable. The amount of misfit in the data was more or less as expected, and no persons (except one who scored full marks) were removed from the data set. A bias analysis was also carried out to look for differences in difficulty for each item in the different groups. Second, several pairs of both native and non-native speakers were trained in introspective reporting on trial cloze tasks, and then asked to complete orally the study task. They were interviewed immediately afterwards about the procedures they had used in attempting to fill the gaps. Third, the answers, errors and corrections made by the 300 genuine candidates were studied to try to understand the sources of difficulty they had actually faced in completing the task. Each of these procedures has provided some insight into the mental processes followed by cloze test candidates. Work is continuing on all three parts concurrently.

A preliminary model of the answering process is given in Figure 1 next page; it will not be discussed in detail here. The Rasch analysis is summarised in Table 1.

Table 1. Analysis of FCE Cloze Test 1990.

Estimated Sample Statistics : Ability Scale	
Mean Ability = 0.847	Standard Deviation = 0.749
Average estimation error	= 0.605
Separation coefficient (sd/se)	= 1.24
Reliability index (like alpha/KR20)	= 0.605

Estimated Sample Statistics : Difficulty Scale	
Mean Difficulty = 0.000	Standard Deviation = 1.356
Average estimation error	= 0.162
Separation coefficient (sd/se)	= 8.39
Reliability index (like alpha/KR20)	= 0.986

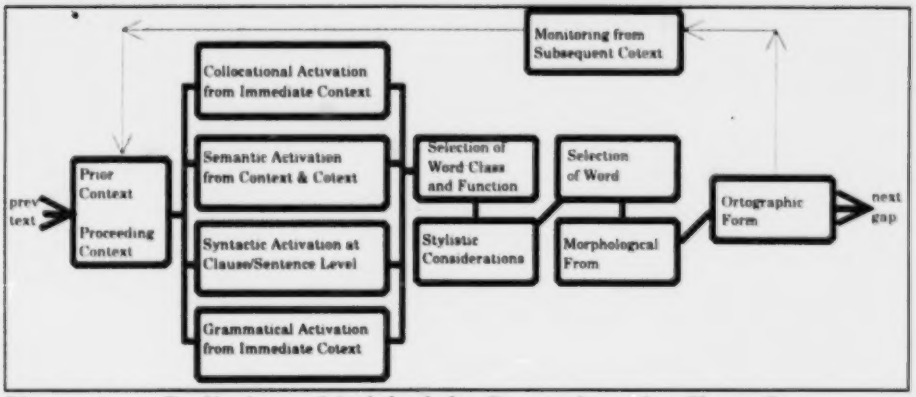


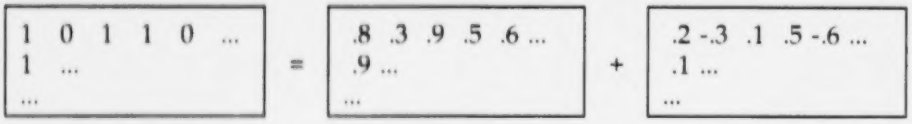
Figure 1. Preliminary Model of the Processing of a Cloze Gap.

It is clear that the question difficulties were successfully separated by the 300 candidates, and that there is sufficient reliable variance in these for a predictive study to have some chance of success. On the other hand, the person abilities were not well separated, with a Rasch reliability coefficient of only 0.605 (the classical KR20 coefficient was 0.64). Stenner's (1983) argument for studying item difficulties is illustrated by these results: the item difficulties were measured with high reliability, to an accuracy almost four times as good as the person abilities.

If item difficulties are to be the source of information about what candidates find difficult, it is important that these difficulties be characteristics of the items rather than the candidates. Before going on to the prediction part of the study, a check was made to see if the question difficulties were stable across L1 groups. This 'bias' study is the main concern of this paper.

2.2. Bias Analysis

A simple Rasch analysis can be seen as extracting a set of parameters measuring persons and items on the first principal component of the data matrix. These parameters can then be used to predict the original data set: each combination of person ability and item difficulty parameters being used to calculate the probability that the person will get that item right. Since these probabilities will never be exactly 0 or 1, there will always be a discrepancy between this predicted observation and the actual one. The **data** matrix is thus decomposed into a **measure**, or model, matrix and a **residual** matrix. The diagram next page illustrates the process.



Data = Measure + Residual

The residual matrix can then be studied in an attempt to understand the nature of any unreliability, or misfit, in the original matrix (Wright, 1989). Within any subset of the residual matrix, the residuals are summed, and this is divided by the measured amount of information in the subset to give a measure of bias in logit units. The equations are:

Person ability	=	b_n	
Item difficulty	=	d_i	
Score of 'n' on 'i'	=	X_{ni}	
Probability that X=1	=	P_{ni} =	$\exp(b_n-d_i) / (1+\exp(b_n-d_i))$
Discrepancy for subset	=	Y_s =	$\Sigma (X_{ni}-P_{ni})$
Information in subset	=	Q_s =	$\Sigma (P_{ni} (1-P_{ni}))$
Bias in subset	=	B_s =	Y_s / Q_s
SE of bias	=	$1/\sqrt{(Q_s)}$	

In this case, the residual matrix was partitioned into 10 sections, corresponding to the 10 countries of origin, and the misfit in each section evaluated for each item in turn. The result is the table of item bias at the end of this article. For convenience of interpretation, the 10 language sets are grouped into three sets: Germanic, Latin and Oriental.

For each item, the difficulty in logits is given on the left; these range from -1.82(easiest) to 2.93(hardest). Under the heading **Bias** is a measure, again in logits, of the size of the apparent bias for that subgroup on that item. This may be understood as the change in difficulty that item seems to have for that L1 group. Thus, for example, on Item 1, we have:

Group	General Difficulty	Bias	Apparent Group Difficulty
Spain	-0.53	-0.96	-1.49
Turks	-0.53	+0.89	+0.36
Japan	-0.53	-0.83	-1.36

The bias **in favour** of Spanish and Japanese students makes the item **easier** for them than for others: the reverse is true for Turks. Alternatively, the Turks may be said to appear, on average, 1.11 logits **less able** on this item than on the rest of the test, and the Spanish and Japanese **more able** on it.

An estimate of the standard error of that group bias is then given, and the ratio of Bias to SE of Bias is reported in the 'Sig' column. This may be interpreted as

a standard normal statistic; values greater than 1.96, the 5% significance level, are marked with an asterisk.

In general, the three language sets in each of the first two groups behaved fairly consistently; the four Oriental sets were, not surprisingly, less consistent, with each of the four groups sometimes seeming rather different from all of the rest.

Although with ten statistics for each item some spurious effects are to be expected, the most significant finding was that only 4 items showed no significant bias effects.

2.3. Understanding the effect

We were surprised how prevalent L1 bias was at the item level. Opinions were sought from native and fluent non-native speakers of the various languages to try to identify interference or transfer effects that might explain these bias effects. The independent reports of typical errors made by learners of English from different language backgrounds in Swan & Smith (1987) were also used. Explanations are suggested below for some of the item effects.

Item 5: Biased in favour of the French and to a lesser extent the other Latins, uniformly somewhat biased against the Orientals. This is a 'pseudo gap', in that no word seems necessary to fill it. It seems natural in French to use phrases like *tous les trains*, whereas we are told that Japanese "only say 60% of what they mean, leaving the reader/hearer to contribute the rest".

Item 6: Favours Dutch and Germans, handicaps Brazilians. 'Connecting trains' is an idiomatic English (and perhaps north European?) phrase. There was a semantic confusion for some candidates — exactly half of the Brazilians seemed to feel that the gap needed a word implying 'need to take' rather than 'miss'.

Item 7: Slightly biased in favour of Dutch and Germans, but handicaps Norwegians and Spanish. 'Odd' behaviour within the Germanic group of languages. According to Swan & Smith, Norwegian would normally place a 'mid-sentence main verb' before an adverb, rather than after as here, and would normally omit the 'us' part of the complement after a 'telling' verb. Thus Norwegians would neither expect the verb 'tell' after an adverb nor expect it before 'us'.

This item seems to show particularly clearly the reasons why cloze may be so prone to L1 bias effects. Given the acknowledged importance of immediate cotext on the response, the intermediate level Norwegian student is faced with a peculiarly difficult task:

a) the immediately preceding word, 'helpfully', suggests that the gap is NOT a verb (perhaps 'announcement' was the verb?).

c) the immediately following word, 'us', suggests that the gap is at least not a *telling* verb, even if it is some other sort of verb.

How then is the unfortunate Norwegian supposed to come up with 'told'? It is the combination of the extremely tight constraints on the answer with the demand that the answer be produced actively rather than recognised passively that handicap the Norwegian on this item rather than others.

All but one of the wrong answers given by Norwegians was 'to', supporting this interpretation. Are Dutch and German more like English in this respect? Some Spanish candidates also wrote 'to', but others gave various answers like 'explained' that would take an indirect rather than a direct object.

Item 9: Strongly favours Dutch and Norwegians, and handicaps most others. This was the most difficult of all the items; even fluent bilingual informants from various language backgrounds (and native children) were defeated by it. Is it purely an adult north European metaphor, to be 'far from' truth?

Item 10: Difficult for Germanic speakers, astonishingly easy for Arabs. This is a fairly straightforward phrasal verb, yet only 11 of the 90 Germanic speakers gave 'about' correctly. 17 Dutch, 12 Germans and 9 Norwegians wrote 'I enquired *for* buses'. A further 26 saw no need for a preposition and wrote some kind of article or quantifier. Perhaps the prepositions in phrasal verbs in English are more tricky than they think. By contrast, we are informed that French speakers (and presumable Arabs) learn phrasal verbs as units.

Item 14: Handicaps French and Brazilians, favours most others. This sort of construction, which involves a noun modifying another noun, seems to be common in most languages, but **not** in Latin ones. French would normally use a 'de' construction, or omit the 'taxi' completely. 14 out of 20 French errors were attempts to insert a verb auxiliary, leaving 'taxi' as the subject of 'warned'.

Item 16: Severely biased against Spanish, favours Japanese. In Spanish, the preposition 'a' can correspond to either 'at' or 'to' in English; of the 30 Spanish speakers 18 chose to write 'arrived at', while 12 chose 'arrived to'.

Item 19: Biased against French, and somewhat against Dutch. The French were simply confused by the inverted structure of the sentence, and offered a variety of ways in which the suitcases would stay — 'here', 'closed', 'still', and so on. As for the Dutch, all but two got the right idea, but 5 misspelled the word 'untill'. (A further 4 corrected their answer by scoring out the extra 'l'). Why are the Dutch so inclined to misspell 'until' with a double 'l'? It may be noted that the British themselves double the 'l' in the abbreviated form 'till' (British teachers seem sat last to have given up insisting that we should write 'til, just as they have given up on 'phone.) Who can blame the Dutch when the English speakers are so illogical?

Item 20: Heavily biased against Japanese, somewhat against Chinese and Turks, and favours French, Spanish and Portuguese speakers. This is perhaps the clearest language group effect of all. Once again we have a 'pseudo gap', where the pronoun is not necessary in English. We are told a Japanese speaker would never feel the need to make explicit such an obvious pronoun, and only one of our Japanese sample did so. Others offered 'money'(6 times), 'fare' (6), 'for' (5), 'it' (3), and several other answers, but no pronouns except for one candidate who offered 'myself'! In contrast, it seems that the Latin languages would naturally include the pronoun as part of the verb structure, the indirect object usually preceding the verb.

3. DISCUSSION

3.1. Conclusion

It looks as if bias with respect to first language may be the rule rather than the exception for cloze items in FL testing.

Some at least of the bias effects observed here are easily explained in terms of L1 transfer or interference. Several researchers have concluded that most cloze items operate at a very local level, with a cotext extending no more than 5 words either way. In a case such as the Norwegians and Item 7, where both the immediately preceding word and the immediately following word would tend to mislead in terms of L1 structure, it is hardly surprising to find a substantial increase in difficulty for Norwegians compared to other groups. From this evidence and from the interview studies, we conclude that cloze is essentially a **productive language task**, in which context, cotext, syntax and grammar combine to set constraints within which the production process must take place. At an intermediate level, where students meet FCE, it must make a considerable difference to their chances if the constraints operate in a way that is familiar from their L1, or if they are forced to work only from a very imperfect knowledge of L2.

3.2. Explanation

The unique features of cloze tests can perhaps be appreciated most clearly by considering cloze in the context of Bialystok's framework for understanding the demands of different language tasks (see, for example, Bialystok, 1984). Because a cloze test is normally written, not oral, and because there is normally enough time for all candidates to complete the test and to revise their answers, the task rates very low on Bialystok's dimension of *Control*. Yet because the candidate has to generate the answer — that is, to generate out of their head the function class, the word class, the word root form, the morphology and finally the orthographic form, using whatever clues the context and cotext offer — the task rates very high on the *Analytic* dimension. It is hard to imagine any language

task that might be set to intermediate learners that would make greater demands on their analysed knowledge of the language. Indeed the cloze task compares quite well in this respect with Bialystok's own extreme example of writing poetry; arguably, poets have more freedom to choose or to avoid any structure they are uncomfortable with, and so need **less** analysed knowledge than the cloze test taker.

3.3. Does Bias Matter?

There are two sets of circumstances in which bias at the item level might not matter: if it is not large compared to errors of measurement, or if it cancels out over the set of items. In this case it is clear that the size of these bias effects cannot be ignored. The bias for or against one or other L1 group often exceeds 1 logit; in 7 of the 20 items it exceeds 1.55 logits for some group, meaning that those candidates appear more than two population standard deviations more or less able on that item than they do on the rest of the test.

Can we assume that these bias effects in the 20 items will cancel each other out overall, for each language group? Or should we suspect that any cloze test is likely to be biased overall in favour of groups whose L1 is lexically and particularly structurally similar to the target language? A preliminary attempt to check this overall effect for these Cambridge data was made by regressing the cloze scores on the scores on the rest of FCE Paper 3. This paper is called 'Use of English', and the other parts, adding up to 52 marks, seem to measure a similar sort of ability. The regression was carried out within each national group, and the slopes varied from 0.241 for the Bahrain sample to 0.999 for the Japanese group. While this variation may be taken as *prima facie* evidence for bias in the test as a whole, it is unconvincing as there was no obvious pattern in the regression slopes. Better evidence came from the graph and table below.

The plot (Figure 2) shows the national group mean ability estimates from the cloze test against those from the rest of Paper 3. There is no simple linear relationship between scores from the two parts of the examination. This shows at least that there is some differential effect between the various groups.

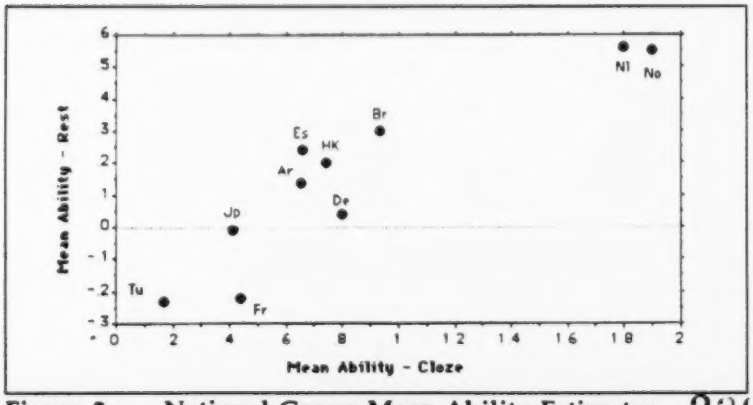


Figure 2. National Group Mean Ability Estimates.

The table contains the results of a single common regression analysis across all groups. It shows the mean standardised residual within each group, and whether or not this mean residual is significant. The Dutch and Norwegians did significantly better on the cloze test than on the rest of the examination, while the Spanish, Japanese and Turks did less well on the cloze. The picture is muddied by the differences in mean proficiency level of the groups, as the Dutch and Norwegian samples were considerably more able than the others. A clearer picture must await the analysis of a sample chosen more deliberately with these hypotheses in mind. Nevertheless, it does look as if the FCE cloze test may be biased in favour of first languages that are similar to English in structure and functional lexis. Such a conclusion would again support the hypothesis that this kind of cloze is essentially a test of productive language competence.

L1 Group	**Mean Standardised Residual**
Dutch (Nl)	+0.75*
Norwegian (No)	+0.90*
Portuguese (Br)	-0.16
Spanish (Es)	-0.35*
Chinese (HK)	-0.18
Arabic (Ar)	-0.23
German (De)	+0.09
Japanese (Ja)	-0.36*
Turkish (Tu)	-0.39*
French (Fr)	-0.39

3.4. Recommendations

Cloze test items do appear to be especially liable to L1 bias. This does not in any way reduce the usefulness of cloze as a teaching technique. In fact it appears that a cloze test designed to highlight the particular difficulties that a particular group may encounter in L2 may be especially useful. Items such as 'I enquired *about* buses', for example, may help Germanic language speakers to realise that English phrasal verbs are not always as simple as they might seem.

Nor do these results reduce the usefulness of cloze as a test in a context of L1 homogeneity. If all of the students share the same L1, then these L1 transfer effects become not a source of bias but a source of evidence about the state of the students' interlanguage.

If the characterisation of cloze as *high analysis, low control* is accurate, we ought to find less overall bias in cloze tests given to more proficient learners, as in Cambridge's CPE, and more bias with lower level learners, as in the Preliminary English Test. This will be looked at in a follow up study.

It is exclusively when the candidates taking a test come from a range of L1 backgrounds, even German+French, or in some cases even German + Norwegian, that L1 transfer effects become a source of bias and so an undesirable part of the trait being measured.

Cloze tests should never be thought of as tests of reading ability.

It does not seem wise for any cloze test in these (perhaps in any circumstances) to test the knowledge of isolated idiomatic phrases.

It seems indefensible not to accept answers which are only misspelled, as in the case of the Dutch and 'until'. It is questionable whether exact morphology should be demanded, given the extra difficulties this seems to set for those languages which handle inflection differently or not at all. These demands have the effect of loading the test heavily with writing skills and so handicapping some groups.

It seems likely that the bias effects will be reduced if a multiple choice variation of cloze is used, since this will reduce the crucial productive and morphological demands, although it is possible that this would make matters worse in some cases, such as in the choice of prepositions for phrasal verbs.

Finally, it seems also that — at least in the first instance — we will have to restrict the sample for our construct validation exercise to a single L1 group, or at least to speakers of one language family. And we will have to include possible L1 interference as a potential source of item difficulty.

REFERENCES

Bachman, L.F. 1990. *Fundamental Considerations in Language Testing.* Oxford University Press.

Bachman, L.F., Davidson, F. & Milanovic, M. 1991. The use of test method characteristics in the content analysis and design of EFL proficiency tests. Paper presented at LTRC, Princeton, NJ. March, 1991.

Bialystok, E. 1984. Strategies in interlanguage learning and performance. In Davies, A., Criper, C. & Howatt, A.P.R. (Eds) *Interlanguage: Proceedings of the seminar in honour of Pit Corder.* Edinburgh University Press.

Brown, G. & Yule, G. 1983. *Discourse Analysis.* Cambridge University Press.

Cronbach, L.J. 1984. *Essentials of Psychological Testing.* New York, Harper & Row.

Embretson, S.E. 1983. Construct validity: Construct representation and nomothetic span. *Psychological Bulletin,* 93, 179-197.

Embretson, S.E. & Wetzel, C.D. 1987. Component latent trait models for paragraph comprehension tests. *Applied Psychological Measurement,* 11, 175-193.

Kintsch, W. 1974. *The Representation of Meaning in Memory.* Hillsdale, NJ, Lawrence Erlbaum.

Kintsch, W. & van Dijk, T.A. 1978. Toward a model of text comprehension and production. *Psychological Review,* 85, 363-394.

Pollitt A.B., Entwistle, N.J., Hutchinson, C.J. & De Luca, C. 1983. *What Makes Exam Questions Difficult?.* Edinburgh, Scottish Academic Press.

Pollitt, A.B. & Hutchinson, C.J. 1986. The validity of reading comprehension tests: What makes questions difficult? In Vincent, D., Pugh, A.K. and Brooks, G. (Eds) *Assessing Reading.* Basingstoke, Macmillan Education.

Stenner, A.J., Smith, M., Burdick, D.S. 1983. Towards a theory of construct definition. *Journal of Educational Measurement,* 20, 305-316.

Swan, M. & Smith, B. 1987. *Learner English.* Cambridge University Press.

Wright, B.D. 1989. *Method and Meaning of Measurement.* Chicago, MESA Laboratory.

Wright, B.D. & Stone, M.H. 1979. *Best Test Design.* Chicago, MESA Press.

APPENDIX 1

FCE Cloze Test: Dec 1990

1 *Fill each of the numbered blanks in the following passage. Use only* ***one*** *word in each space.*

One of the worst journeys I have ever experienced occurred a few weeks ago. I (1) ..... booked a cheap flight to Switzerland so the ticket could not be changed (2) ..... any way. (3) ..... I missed the flight, I would not be (4) ..... to use the ticket for any alternative journey.

When I reached the railway station, I was told that (5) ..... the trains were running late; this meant I would (6) ..... the connecting train for the airport. A loudspeaker announcement helpfully (7) ..... us that an extra train would (8) ..... provided, so we need not worry. Nothing could have been further from the (9) ..... . The extra train did not arrive. I inquired (10) ..... buses, but the last one for the airport had left!

There was only one solution; I (11) ..... to take a taxi. Dragging my suitcases behind me I hurried outside (12) ..... found a taxi. "It's a long way. It'll (13) ..... you a lot", the taxi (14) ..... warned me. I knew that, but the taxi fare would be cheaper (15) ..... having to buy another plane ticket.

We arrived (16) ..... the airport with about twenty minutes (17) ..... spare. I jumped out, looked in my bag (18) ..... my purse and to my horror discovered it was missing! The taxi driver was pleasant but firm. My suitcases were locked in the boot and there they would stay (19) ..... I found a way of paying (20) ..... . Of course, I never caught my plane.

Answers: **Must be one of the following words.**

1 had 2 in 3 If, Assuming 4 able, allowed, permitted,
5 all, unfortunately, regrettably 6 miss (~~lose~~) entitled (~~authorised~~) (~~today~~)
7 told, informed, assured, advised 8 be
9 truth (~~facts, reality, situation, case~~) 10 about (~~into~~)
11 had, decided, needed and 13 cost 14 driver, -driver
15 than 16 at, in 17 to 18 for 19 until, till, unless 20 him, up

APPENDIX 2

*** Item Bias by L1 in FCE Cloze ***

Note that a positive bias (and positive Sig entry) means that the item is harder for that group than expected, that is, they did less well than expected.

Bias in each Group on Item 1		Gp	N	Bias	SE Bias	Noise	Sig
d = -0.53	Netherlands	1	30	-0.24	0.59	0.73	-0.41
	Germany	2	30	-0.22	0.45	0.77	-0.48
	Norway	3	29	-0.02	0.63	0.85	-0.03
	Spain	4	30	-0.96	0.44	0.61	-2.16*
	Brazil	5	30	-0.64	0.46	0.75	-1.38
	France	6	30	+0.37	0.42	1.10	+0.88
	Turkey	7	30	+0.89	0.39	1.17	+2.26*
	Japan	8	30	-0.83	0.41	0.71	-2.00*
	Hong Kong	9	30	+0.14	0.45	0.86	+0.30
	Bahrain/Oman	10	30	+0.64	0.44	1.32	+1.47

Bias in each Group on Item 2		Gp	N	Bias	SE Bias	Noise	Sig
d = 0.21	Netherlands	1	30	+0.16	0.48	1.32	+0.33
	Germany	2	30	-0.35	0.40	0.96	-0.88
	Norway	3	29	-0.10	0.50	1.13	-0.21
	Spain	4	30	+0.46	0.40	1.13	+1.15
	Brazil	5	30	+0.02	0.42	0.91	+0.04
	France	6	30	-0.08	0.39	1.12	-0.21
	Turkey	7	30	-0.47	0.38	1.09	-1.25
	Japan	8	30	+0.35	0.38	1.01	+0.92
	Hong Kong	9	30	-0.40	0.40	1.20	-0.99
	Bahrain/Oman	10	30	+0.15	0.39	1.05	+0.38

Bias in each Group on Item 3		Gp	N	Bias	SE Bias	Noise	Sig
d = 0.08	Netherlands	1	30	+1.30	0.50	1.45	+2.60*
	Germany	2	30	-0.06	0.41	0.98	-0.15
	Norway	3	29	+0.57	0.52	1.35	+1.10
	Spain	4	30	+0.12	0.40	0.93	+0.29
	Brazil	5	30	-0.20	0.42	1.04	-0.48
	France	6	30	-0.10	0.39	0.98	-0.25
	Turkey	7	30	-0.62	0.38	0.98	-1.65
	Japan	8	30	-0.54	0.38	0.91	-1.41
	Hong Kong	9	30	+0.39	0.41	1.25	+0.96
	Bahrain/Oman	10	30	-0.34	0.40	0.87	-0.85

Bias in each Group on Item 4		Gp	N	Bias	SE Bias	Noise	Sig
d = -1.82	Netherlands	1	30	+0.67	0.94	1.27	+0.71
	Germany	2	30	-0.25	0.67	0.90	-0.37
	Norway	3	29	+0.03	1.05	1.01	+0.03
	Spain	4	30	-0.82	0.63	0.48	-1.29
	Brazil	5	30	+0.19	0.67	1.16	+0.29
	France	6	30	+0.93	0.59	1.59	+1.57
	Turkey	7	30	-0.27	0.55	0.89	-0.48
	Japan	8	30	-0.10	0.60	0.86	-0.16
	Hong Kong	9	30	-0.31	0.66	0.76	-0.48
	Bahrain/Oman	10	30	-0.73	0.65	0.53	-1.13

Bias in each Group on Item 5		Gp	N	Bias	SE Bias	Noise	Sig
d = -0.20	Netherlands	1	30	-0.21	0.54	1.14	-0.40
	Germany	2	30	-0.50	0.42	0.76	-1.17
	Norway	3	29	+0.65	0.57	1.65	+1.15
	Spain	4	30	-0.81	0.42	0.64	-1.94
	Brazil	5	30	-0.69	0.44	0.84	-1.58
	France	6	30	-1.09	0.40	0.73	-2.74
	Turkey	7	30	+0.37	0.38	0.93	+0.97
	Japan	8	30	+0.64	0.39	1.08	+1.64
	Hong Kong	9	30	+0.71	0.42	1.22	+1.68
	Bahrain/Oman	10	30	+0.60	0.41	1.04	+1.46

BEST COPY AVAILABLE

Bias in each Group on Item 6		Gp	N	Bias	SE Bias	Noise	Sig
d = 0.78	Netherlands	1	30	-1.14	0.44	0.73	-2.63*
	Germany	2	30	-1.19	0.39	1.05	-3.06*
	Norway	3	29	+0.01	0.45	0.91	+0.03
	Spain	4	30	+0.34	0.39	0.90	+0.88
	Brazil	5	30	+0.94	0.41	0.72	+2.32*
	France	6	30	-0.51	0.39	1.04	-1.29
	Turkey	7	30	+0.74	0.39	0.73	+1.88
	Japan	8	30	+0.07	0.38	0.99	+0.19
	Hong Kong	9	30	-0.02	0.39	0.93	-0.04
	Bahrain/Oman	10	30	+0.47	0.38	0.83	+1.22

Bias in each Group on Item 7		Gp	N	Bias	SE Bias	Noise	Sig
d = 0.28	Netherlands	1	30	-0.82	0.48	0.52	-1.72
	Germany	2	30	-0.73	0.40	0.94	-1.83
	Norway	3	29	+1.31	0.50	1.88	+2.65*
	Spain	4	30	+1.02	0.40	1.12	+2.57*
	Brazil	5	30	-0.05	0.41	0.98	-0.11
	France	6	30	+0.01	0.39	0.88	+0.01
	Turkey	7	30	-0.54	0.38	1.09	-1.43
	Japan	8	30	-0.44	0.38	1.00	-1.15
	Hong Kong	9	30	+0.02	0.40	1.01	+0.06
	Bahrain/Oman	10	30	+0.24	0.39	1.11	+0.61

Bias in each Group on Item 8		Gp	N	Bias	SE Bias	Noise	Sig
d = -1.96	Netherlands	1	30	+0.90	1.00	2.01	+0.90
	Germany	2	30	+1.89	0.71	2.54	+2.66*
	Norway	3	29	-1.06	1.12	0.06	-0.95
	Spain	4	30	+0.13	0.67	1.31	+0.20
	Brazil	5	30	-0.64	0.71	0.63	-0.90
	France	6	30	-0.76	0.62	0.46	-1.22
	Turkey	7	30	+0.53	0.58	1.43	+0.91
	Japan	8	30	-1.14	0.63	0.14	-1.82
	Hong Kong	9	30	+0.28	0.69	1.04	+0.41
	Bahrain/Oman	10	30	-1.13	0.68	0.13	-1.66

Bias in each Group on Item 9		Gp	N	Bias	SE Bias	Noise	Sig
d = 2.93	Netherlands	1	30	-1.66	0.44	1.33	-3.75*
	Germany	2	30	+0.28	0.57	0.73	+0.49
	Norway	3	29	-1.43	0.44	1.30	-3.29*
	Spain	4	30	-0.51	0.59	1.23	-0.86
	Brazil	5	30	+1.11	0.54	0.42	+2.06*
	France	6	30	+1.16	0.64	0.27	+1.81
	Turkey	7	30	+1.09	0.74	0.10	+1.48
	Japan	8	30	+1.11	0.67	0.12	+1.67
	Hong Kong	9	30	+1.19	0.57	0.19	+2.07*
	Bahrain/Oman	10	30	+1.15	0.60	0.15	+1.92

Bias in each Group on Item 10		Gp	N	Bias	SE Bias	Noise	Sig
d = 2.32	Netherlands	1	30	+1.52	0.41	0.99	+3.67*
	Germany	2	30	+1.17	0.48	0.34	+2.45*
	Norway	3	29	+0.64	0.41	0.99	+1.58
	Spain	4	30	-0.33	0.49	1.31	-0.67
	Brazil	5	30	-0.84	0.47	1.53	-1.78
	France	6	30	-0.11	0.53	1.11	-0.21
	Turkey	7	30	+0.49	0.58	0.62	+0.83
	Japan	8	30	+0.07	0.53	0.84	+0.13
	Hong Kong	9	30	-0.03	0.48	0.98	-0.06
	Bahrain/Oman	10	30	-3.08	0.49	2.79	-6.27*

Bias in each Group on Item 11		Gp	N	Bias	SE Bias	Noise	Sig
d = -0.86	Netherlands	1	30	-0.37	0.66	0.83	-0.56
	Germany	2	30	-0.14	0.49	0.73	-0.28
	Norway	3	29	-0.66	0.71	0.56	-0.93
	Spain	4	30	-0.07	0.48	0.82	-0.15
	Brazil	5	30	+0.11	0.50	0.81	+0.22
	France	6	30	-0.62	0.45	0.61	-1.39
	Turkey	7	30	-0.39	0.42	0.83	-0.93
	Japan	8	30	+0.19	0.44	0.88	+0.43
	Hong Kong	9	30	+0.76	0.49	1.28	+1.55
	Bahrain/Oman	10	30	+0.22	0.47	1.11	+0.46

Bias in each Group on Item 12		Gp	N	Bias	SE Bias	Noise	Sig
d = 0.32	Netherlands	1	30	-0.18	0.47	0.78	-0.38
	Germany	2	30	+0.02	0.40	0.83	+0.04
	Norway	3	29	-0.69	0.49	0.75	-1.40
	Spain	4	30	+0.50	0.40	1.05	+1.26
	Brazil	5	30	+0.58	0.41	0.95	+1.42
	France	6	30	+0.55	0.39	1.15	+1.44
	Turkey	7	30	-0.30	0.38	0.93	-0.80
	Japan	8	30	-0.20	0.38	0.79	-0.52
	Hong Kong	9	30	-0.82	0.40	0.75	-2.05*
	Bahrain/Oman	10	30	+0.04	0.39	1.01	+0.09

Bias in each Group on Item 13		Gp	N	Bias	SE Bias	Noise	Sig
d = -0.67	Netherlands	1	30	-0.11	0.62	0.76	-0.18
	Germany	2	30	-0.52	0.47	0.63	-1.10
	Norway	3	29	-1.18	0.66	0.18	-1.79
	Spain	4	30	+0.59	0.46	1.13	+1.30
	Brazil	5	30	+0.38	0.48	1.12	+0.80
	France	6	30	-0.75	0.43	0.75	-1.74
	Turkey	7	30	+0.27	0.40	1.07	+0.68
	Japan	8	30	+0.36	0.43	1.13	+0.85
	Hong Kong	9	30	+0.08	0.47	1.11	+0.17
	Bahrain/Oman	10	30	-0.38	0.45	0.90	-0.85

Bias in each Group on Item 14		Gp	N	Bias	SE Bias	Noise	Sig
d = -1.23	Netherlands	1	30	-0.61	0.75	0.63	-0.82
	Germany	2	30	-0.95	0.55	0.27	-1.73
	Norway	3	29	-1.11	0.82	0.11	-1.35
	Spain	4	30	-0.50	0.53	0.68	-0.95
	Brazil	5	30	+1.16	0.55	1.75	+2.11*
	France	6	30	+3.26	0.49	3.03	+6.66*
	Turkey	7	30	-1.09	0.46	0.39	-2.39*
	Japan	8	30	-0.32	0.49	0.77	-0.64
	Hong Kong	9	30	-1.28	0.54	0.28	-2.37*
	Bahrain/Oman	10	30	-0.42	0.53	0.85	-0.79

Bias in each Group on Item 15		Gp	N	Bias	SE Bias	Noise	Sig
d = -1.77	Netherlands	1	30	+3.20	0.93	3.61	+3.44
	Germany	2	30	+0.59	0.66	1.58	+0.89
	Norway	3	29	+1.06	1.03	2.07	+1.03
	Spain	4	30	+0.72	0.62	1.28	+1.16
	Brazil	5	30	-0.29	0.66	0.70	-0.44
	France	6	30	+0.18	0.58	1.16	+0.31
	Turkey	7	30	-1.18	0.54	0.18	-2.18*
	Japan	8	30	-1.17	0.59	0.17	-1.99*
	Hong Kong	9	30	-0.35	0.65	0.75	-0.53
	Bahrain/Oman	10	30	-0.75	0.63	0.51	-1.18

Bias in each Group on Item 16		Gp	N	Bias	SE Bias	Noise	Sig
d = -1.32	Netherlands	1	30	+0.04	0.78	0.89	+0.06
	Germany	2	30	-0.91	0.57	0.26	-1.62
	Norway	3	29	-0.38	0.85	0.66	-0.45
	Spain	4	30	+2.20	0.54	2.88	+4.07*
	Brazil	5	30	-0.28	0.57	0.98	-0.50
	France	6	30	-0.26	0.50	0.83	-0.51
	Turkey	7	30	-0.40	0.47	0.78	-0.85
	Japan	8	30	-1.26	0.50	0.25	-2.49*
	Hong Kong	9	30	-0.33	0.56	1.00	-0.60
	Bahrain/Oman	10	30	+0.83	0.54	1.36	+1.53

Bias in each Group on Item 17		Gp	N	Bias	SE Bias	Noise	Sig
d = 2.56	Netherlands	1	30	-0.98	0.42	1.32	-2.33*
	Germany	2	30	+1.06	0.51	0.49	+2.10*
	Norway	3	29	-0.26	0.41	0.99	-0.63
	Spain	4	30	+0.76	0.53	0.59	+1.43
	Brazil	5	30	+0.05	0.50	0.96	+0.10
	France	6	30	+0.26	0.57	0.63	+0.45
	Turkey	7	30	+1.13	0.64	0.14	+1.77
	Japan	8	30	+0.49	0.58	0.62	+0.85
	Hong Kong	9	30	-2.10	0.51	2.21	-4.13*
	Bahrain/Oman	10	30	+1.22	0.53	0.22	+2.30*

Bias in each Group on Item 18		Gp	N	Bias	SE Bias	Noise	Sig
d = 0.24	Netherlands	1	30	-0.11	0.48	1.03	-0.22
	Germany	2	30	+1.37	0.40	1.33	+3.43*
	Norway	3	29	+0.12	0.50	0.74	+0.23
	Spain	4	30	-0.52	0.40	1.01	-1.32
	Brazil	5	30	-0.01	0.41	0.92	-0.03
	France	6	30	-1.01	0.39	1.08	-2.61*
	Turkey	7	30	-0.08	0.38	1.05	-0.21
	Japan	8	30	-0.41	0.38	1.00	-1.07
	Hong Kong	9	30	+0.06	0.40	0.87	+0.14
	Bahrain/Oman	10	30	+0.42	0.39	1.03	+1.08

Bias in each Group on Item 19		Gp	N	Bias	SE Bias	Noise	Sig
d = -0.25	Netherlands	1	30	+0.72	0.54	1.34	+1.33
	Germany	2	30	-0.45	0.43	0.93	-1.05
	Norway	3	29	-0.60	0.58	0.40	-1.04
	Spain	4	30	-0.59	0.42	0.69	-1.41
	Brazil	5	30	+0.07	0.44	0.91	+0.15
	France	6	30	+0.88	0.40	1.21	+2.19*
	Turkey	7	30	-0.16	0.38	0.84	-0.41
	Japan	8	30	+0.24	0.40	1.14	+0.60
	Hong Kong	9	30	-0.32	0.43	0.75	-0.74
	Bahrain/Oman	10	30	-0.19	0.41	0.87	-0.47

Bias in each Group on Item 20		Gp	N	Bias	SE Bias	Noise	Sig
d = 0.89	Netherlands	1	30	+0.07	0.43	1.17	+0.17
	Germany	2	30	+0.52	0.39	0.85	+1.33
	Norway	3	29	+0.09	0.44	0.98	+0.22
	Spain	4	30	-1.62	0.39	1.19	-4.12*
	Brazil	5	30	-1.31	0.41	1.04	-3.23*
	France	6	30	-0.78	0.40	1.01	-1.98*
	Turkey	7	30	+0.97	0.40	0.68	+2.43*
	Japan	8	30	+1.62	0.39	0.78	+4.17*
	Hong Kong	9	30	+0.94	0.39	1.03	+2.41*
	Bahrain/Oman	10	30	-0.68	0.38	1.11	-1.75

PROFICIENCY TESTS IN CONTEMPORARY FOREIGN LANGUAGE SYLLABUS

D. Rumpite, A. Rotanova and L. Ilyinskaya
Riga Technical University, Latvia

The article presents a project in which proficiency tests were developed for modern languages. The project was based on Vygotsky's cultural-historical theory of learning and development, which has received considerable attention also in the west in recent years. The testing system attempts to take account of the genesis and essential structure of language skills, the role of feedback and of self-evaluation. Empirical results obtained in the project conclude the article.

INTRODUCTION

Historical periods in foreign language teaching methodology are characterized by certain dominating approaches and linguistic, psychological and educational theories. But there has always been one central issue — how to facilitate and accelerate language learning bringing it in line with the needs and level of the society. Vygotsky's cultural-historical theory of the integral unity of the psychic and material processes and activities within a socio-cultural context was used as the basis of our approach. It implies that a student learns how to be socially appropriate in a language performance in a variety of socio-cultural situations and this ability characterizes the attained language and communicative proficiency.

Linguistic phenomena are selected according to their priority in actual communication which allows the learners to interact within a wide range of communicative language functions appropriate to a certain cultural setting.

In specifying the foreign language content an attempt was made to identify the fundamental structural units of a particular subject matter. Students learn the fundamental "rules" rather than numerous unrelated separate objects. They refer to all the language skills including grammatical competence, reading strategies, and speaking. Thus the tests become more objective, purposeful and considerably fewer as they correspond to the most essential ("invariant") and underlying concepts, tasks and situations which are necessary to acquire a particular skill.

The tests also reflect the genesis of the skill formation and its different levels, the operational content of the activity.

Using the above approach we have tried to introduce tests with a feedback potential — meaningful multiple choice tasks, interviews, dialogue games, a combination of audio and video material with a graphic (or written) presentation of an answer taking into account the goals set and the language and communicative proficiency level of the students.

The results of the research confirm that if students themselves take part in the process of testing and self-evaluation in accordance with the set goals and definite criteria known to the student the adequacy of result evaluation and self evaluation increases and the cognitive activity level of the student grows.

In the Baltic countries – Latvia, Estonia, Lithuania – special attention is being paid now to mastering foreign languages because the political status of these countries has opened up new prospects for expanding economic, scientific, cultural and other ways of cooperation with the world community. Thus the knowledge of foreign languages has become a practical necessity and even a demand for most of the specialists in various fields. That is why one of the primary tasks of the education reform taking place now is to improve the language teaching efficiency, introducing more up-to-date and effective technologies which would to a great extent guarantee the expected result. An objective difficulty is the absence of foreign language environment and rather uncertain possibilities to stay or study abroad, which would greatly facilitate reaching proficiency in a foreign language.

The new social environment has radically changed motivation, the goals and the content of instruction. That is why valid and reliable tests as an estimate of the cooperative learning and interaction between a teacher and a learner reflecting these new tendencies are of particular value and interest.

1. TEST DESIGN: A RATIONALE

Each test serves its specific purpose and is based in fact on some definite theoretical approach (in psychology, education theory and respective subject fields). In elaborating tests we proceed from the following assumptions:

1) There is a direct correlation between the model of teaching, learning and testing. Tests embody the principal features of training while the latter determines the nature of tests. Now learning becomes more personal, student centered, with social and philosophical values gaining more personal relevance. In this connection Vygotsky's cultural-historical theory of the unity of psychic and material processes and activities served as the basis of our approach. According to Vygotsky, psychic and cognitive activities are considered in close

association with man's practical actions and interrelations with the surrounding world. The major element in the development of human psyche is the assimilation of social experience. The ability of students to be socially appropriate in their use of the language is an important part of their linguistic and communicative competence. Vocabulary, grammar and culture phenomena are selected and included in the syllabus according to their priority in actual communication and the goals of studying a language. Personality development is attained through critical thinking and cooperative activities.

Though tests preserve their basic features of objectivity, reliability and validity, they become more complex, elaborate and at the same time more natural as the final object of testing is the specific communicative competence, the linguistic competence being only an integral part of it.

When testing we have in mind the following interconnected components of a foreign language syllabus: goals and objectives, language content, process and product.

2) The goals of the course is the most important criterion to check the level in skill development. The goals comprise the communicative element (what skills are needed to maintain interpersonal relations: exchange of information, ideas, opinions on a certain subject matter), socii-cultural element (social rules, roles and interactions) as well as an operational element (what actions should be learnt to acquire the skill).

As to the language content, the key question is what we test and what is behind the test: do we test separate elements or interconnected parts of the acquired language system in its practical use. If the content of the subject does not give adequate orientation in a language since the orientation is not on essential linguistic features but on scattered elements of the studied language, the student has to memorize all the particulars and can then function within these particulars, but not with the whole system of language phenomena. To master the content of such a language course, mostly memory and much less thinking is needed. In this case mastering of a foreign language is performed through the assimilation of structure patterns by frequent mechanical repetition. Then the actions of the students and the exercises which they complete in order to master any part of the programme for the most part are not appropriate to the goal and are often contrary to it. Presumably the test scores will also be low.

We follow the assumption that any action includes three parts: orienting, execution and control. The orienting component of the action is concerned with correct and rational formulation of the executive part.

Research findings have shown that the orienting component plays a decisive role in the development of actions. It determines both the rate at which the actions develop, their quality and, accordingly, the outcome demonstrated in the tests.

We are also aware of the object of testing — whether it relates to individual operations or to their final outcome. We observe a positive influence when testing is linked with individual operations. If it is carried out only in relation to the final result it exerts no perceptible influence on the quality of assimilation. During the initial stages of the process of assimilation testing relates to individual operations.

The same refers to student activities included in the so-called "process syllabus". An important precondition is noted here: the activity (skill) to be developed needs to be structurally analysed as to the hierarchy of its separate operations to estimate the gradual controlled process of skill development. Such a system of tests reflects the genesis of skill formation and its particular levels.

In specifying the foreign language content one of the trends (P.Galperin, O.Kabanova) is to point out the underlying, fundamental units of the subject matter. When basic structural units have been identified the full diversity of particular situations can be derived. In such a context the language phenomena perform a new function: they serve as a means for assimilating the essence rather than as objects of assimilation (N. Talyshina).

2. TESTING READING COMPREHENSION

One of the basic aims in mastering a foreign language at a higher technical school is to achieve an appropriate level of reading skills to show whether the information contained in the text has been fully comprehended. In our case not only the formal content mattered, but particularly the meaning and the message it communicates.

Each test was aimed at checking a certain level of comprehension (overall meaning, particular facts, most important facts, general idea) corresponding to the types of reading – scanning, skimming, study reading.

The test covered the three main and interconnected actions that make up the process of reading, expected to have been mastered, namely, 1) decoding of the semantic level of the text; 2) deriving the factual information; 3) adequate understanding and interpretation of its lexical and grammatical (linguistic) level.

As the content of the reading process was determined in accordance with the basic characteristics at different levels of formation and geared to the final and intermediate goals, the tests were assumed to be adequate to their purpose. They included multiple choice items dealing with the whole text, separate passages, sentences within the context of the passage, or with individual words within a certain context, as well as graphical charts (Table 1).

Table 1. Operational Content of Reading Activity and Corresponding Tests.

Object of testing – reading comprehension	The tests used
1. Ability to draw conclusions, to evaluate facts and to get the meaning and the message of the text	Multiple choice tasks on the level of the whole text and separate passages ("Which statement best expresses the main idea"; "All the following ideas appear in the passage. Which do you think the author attributes most importance to"; "Which would be the most suitable title to the text"). Graphical charts representing the relationship between the main idea and subordinate details. Multiple choice task of anticipation.
2. Ability to compare parts of the texts, to summarize facts and to identify separate facts. Ability to comprehend a sentence and a passage in their integrity	Multiple choice tasks on the level of separate passages and sentences ("Which of the following facts are discussed in the passage"; "Which passage supports the following conclusion"; "Identify logical relationships between the passages"; "Choose the most appropriate ending for each incomplete sentence"). Graphs on the logical structure of the text.
3. Ability to infer the meaning of a word within a sentence	Task "Fill in (or: choose) the appropriate word in a sentence"; Task on contextual reference.

It may be concluded that a test will have fulfilled its diagnostic function if not only the final result is aimed at and attained, but also the intermediate levels are taken into account. So we may have a more complete and objective view of the level of assessment and proficiency.

Thus it is possible to reduce the number of exercises considered necessary to acquire the skill as well as the number of test items, and at the same time to have higher scores (Figure 1).

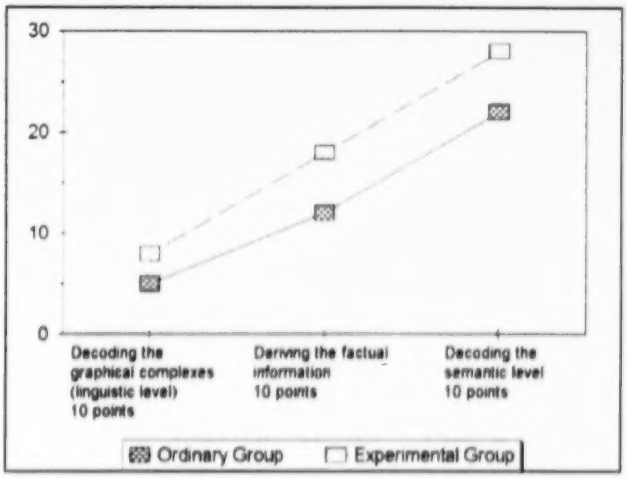


Figure 1. Comparative Results in Testing Reading, Comprehension in Experimental and Ordinary Groups.

3. TESTING ORAL COMMUNICATION SKILLS

Another important task is to teach students to communicate, to express their ideas, opinions, and share feelings in a certain situation. The tests compiled to test the communicative skills of students in a foreign language check their ability 1) to comprehend the situation and their social role in it; 2) to choose the appropriate linguistic means to convey the message; 3) to express themselves; and 4) to understand the response.

We give oral as well as written tests and give discrete point tests as well as integrative tests (testing the communicative competence, interaction, the appropriateness of what is said, i.e., testing all the items together in a realistic situation).

We made an attempt to classify the numerous communicative language functions and identify the basic, underlying ones from which the others may be derived. The latter were included in the syllabus. They are connected with 1) explaining something ("direction", "intention", "possibility" etc.); 2) convincing ("necessity", "obligation", "certainty" etc.); 3) approving ("praise", "likes", "agreement", "preference", "pleasure" etc.); 4) disapproving ("dislike", "disagreement" etc.). The acquired language functions served as one of the means to express one's attitude towards an action or a fact.

In taking the tests the students had to 1) respond to the verbal stimulus which corresponded to the most typical socii-cultural settings; 2) take part in an

interview relating to various situations; and 3) perform a certain social role following the scheme presented on the role card (in a role play).

Written communicative tests were also used. They included the following tasks "Match the stimuli and possible responses", "Select the unrelated word from a group of words", "Match the communicative expressions which express the same function"; "Study the situation. Tell whether the language is appropriate or inappropriate by circling "True" or "False"; "Study the situation and choose the appropriate response".

The content of the tests and the degree of their complexity depended on the level of a particular foreign language course. Thus, there is an interrelated set of tests to correspond to the obligatory and optional language courses at the university.

When the factors in the situation require it, students' skill in encoding and decoding is checked in integrated listening, speaking, reading and writing tasks. Audio-visual material is used here. Students watch such video films as "Britain", "London", "American President's Inauguration Ceremony", "Computers" and perform the following test: 1) Make authentic responses; 2) Do a multiple choice task; 3) Formulate questions; 4) Answer questions; 5) Make a short presentation.

Results obtained in developing and testing different communicative skills show once again that test scores depend very much on the way how teaching and learning are organized. Students with initially lower scores have better assessment results if the content and structure of the communicative activity have been made explicit and comprehended.

4. SELF-EVALUATION

Another aspect of foreign language testing may be mentioned. It is rather evaluation and self-evaluation of students' learning-cognitive activity results, or more precisely, the adequacy of evaluation and self-evaluation.

Related to the self-evaluation aspect of testing, it might be necessary to point out that the evaluation stage is only the final stage of the students' learning--cognitive activity process in which three stages are traditionally distinguished: 1) setting of a goal, 2) realization and 3) results. In fact all the three stages are closely interrelated and it is practically impossible to speak about the result evaluation stage without mentioning the previous two as, on the other hand, it is necessary to evaluate the results in relation to the goal to be achieved, the contents chosen or the methods applied. On the other hand, adequate evaluation and self-evaluation of the results facilitates setting a new goal and makes the possibility of setting this goal more objective.

Thus, in the first stage, students are taught to set a concrete, either a professionally or generally, but that is most important — a personally significant goal for acquiring a foreign language. They set the goal either with the help of the instructor or independently. The personal significance of the goal made the students highly motivated.

In the second stage, students either choose the means for acquiring their goal ((the contents and methods) from those offered by the instructor, or they determine the means themselves.

In the third stage, students' activity results are evaluated by the instructor, fellow students and the student him/herself and the three evaluations are compared.

We proceed from the assumption that the setting of a goal manifests itself in understanding of the goal by students, realization — in the emotional attractiveness of the acquiring of a foreign language as well as in the students' freedom of choice and creative seeking of means, and the results — in the adequacy of evaluation and self-evaluation.

It was necessary to measure five components or manifestations of the three learning-cognitive activity stages mentioned above. That is why the criteria of these manifestations were worked out both in objective and subjective terms. Objective criteria were mostly those that the instructor could observe objectively or measure. Subjective criteria are mostly connected with the students' attitudes. Of course, this division was only relative, as the students' subjective attitude was also objectively measured. On the basis of these criteria, the streaming of the experimental group students was done and three groups were distinguished for cognitive activity as a whole and for each of the cognitive activity components separately: 1) low; 2) intermediate and 3) high.

Some of the adequacy of evaluation and self-evaluation criteria and manifestations were as follows:

1) low level — the evaluation is not adequate to self-evaluation. The student is not able to give adequate evaluation of his activity results in agreement to the goal to be achieved. His self-evaluation is either too high or too low. The instructor's assistance is necessary for adequate evaluation;
2) intermediate level — the evaluation is partly adequate to self-evaluation. The student is able to partly adequately evaluate the results of his activity. Still, for adequate self-evaluation and evaluation an instructor's help may be necessary;
3) high level — self-evaluation is adequate to evaluation. The student self-evaluates his activity results adequately and in accordance with the goal to be achieved.

There were several principles that were observed in the result evaluation process. First of all, as it was already mentioned above, the results were evaluated and the skills were tested according to the goal to be achieved.

Secondly, the results were evaluated by the students themselves (self-evaluation), by the fellow students and by the instructor according to definite criteria worked out by the instructor with the active participation of the students.

Thirdly, the students progress was tested and their activity results were evaluated during every class as well as at the end of a longer period — a term, a year ,etc. Thus we distinguished a short-period and a long-period evaluation.

It might be of interest to note that in the result evaluation process the so-called criterion of communicative sufficiency was used, according to which a mistake that does not cause misunderstanding is not considered to be a mistake. This was important in creating communicative competence of the students.

The experimental work was conducted in two periods, or we may also call them stages. The manifestations of the learning-cognitive activity components of the experimental group students before (Figure 2), after the first (Figure 3) and the second stage (Figure 4) of the experiment can be compared.

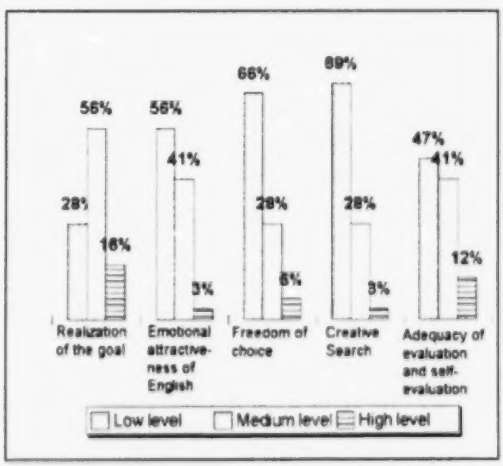


Figure 2. Levels of Cognitive Activity: Experimental groups before the experiment.

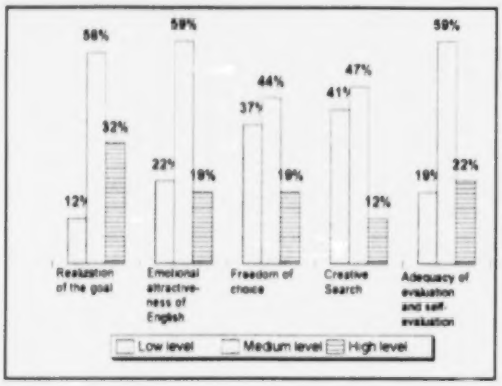


Figure 3. **Levels of Cognitive Activity: Experimental groups after the first stage of the experiment.**

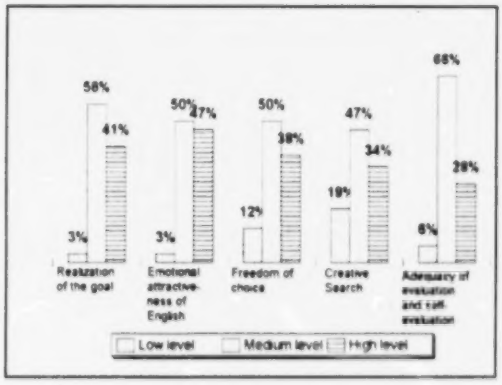


Figure 4. **Levels of Cognitive Activity: Experimental groups after the second stage of the experiment.**

The results of the investigation showed that if students take an active part in their learning-cognitive activity process as a whole and in the results evaluation and self-evaluation in particular in accordance with the goal to be achieved and the definite criteria known to the students the adequacy of the result evaluation and self-evaluation increases and the cognitive activity of the students grows.

REFERENCES

Bachman, L., and Palmer, A. 1981. The construct validation of the FSI oral interview. *Language Learning* 31.(67-86).

Canale, M. and Swain, M. 1980. Theoretical basis of communicative approaches to second language teaching and testing. *Applied Linguistics* 1. (1-47).

Kabanova, V. 1988. The teaching of foreign languages. *Instructional Science* (1-47).

Stansfield, C. & Kenyon, D. & Paiva, R. 1990. The development and validation of the Portuguese speaking test. *Hispania* 73. 3. (642-651) .

Talysina, N. 1981. *The psychology of learning*. Moscow.

Weir, C. 1990. *Communicative language testing*. Prentice Hall.

Authors' Addresses

Vivien Berry
805 C Mansion
Yamake-dori 1-28
Showa-ku
Nagoya 466
JAPAN

Fred Davidson
Division of English as an
International Language
University of Illinois
707 S. Mathews, Rm. 3070
Urbana, Illinois 61801
USA

Harry Gradman and Edith Hanania
Center for English Language
Training
Indiana University
Memorial Hall 313
Bloomington, Indiana 47405
USA

Ari Huhta
Language Centre for Finnish
Universities
University of Jyväskylä
P.O.Box 35
40351 Jyväskylä
FINLAND

Christo Kaftandjiev
Sofia University
BULGARIA

Dorry Mann Kenyon
and Charles W. Stansfield
Center for Applied Linguistics
1118 22nd Street, NW
Washington, D.C. 20037
USA

Dr. Rainer H. Lehmann
Universität Hamburg
Fachbereich Erziehungswissenschaft
Von-Melle-Park 8
2000 HAMBURG 13
DEUTSCHLAND

Brian Lynch
Applied Linguistics
3300 Rolfe Hall, UCLA
Los Angeles, California 90024
USA

Kyle Perkins
and Sheila R. Brutten
Department of Linguistics
Southern Illinois University
Carbondale, Illinois 62901-4517
USA

Alastair Pollitt and Lynda Taylor
Research Centre for English and
Applied Linguistics
Keynes House
Trampington Street
Cambridge CB2 1QA
GREAT BRITAIN

Diana Rumpite, Anna Rotanova
and L. Ilyanskaya
Foreign Languages Department
Riga Technical University
1 Kalku Str.
226355 Riga
LATVIA

Kari Sajavaara
Department of English
University of Jyväskylä
P.O.Box 35
40351 Jyväskylä
FINLAND

Elana Shohamy
Tel Aviv University
School of Education
Tel Aviv
ISRAEL 69978

Felly Stoyanova
Sofia University
BULGARIA

Sauli Takala
Institute for Educational Research
University of Jyväskylä
P.O.Box 35
40351 Jyväskylä
FINLAND

Ingrid Wijgh
Cito
Nieuwe Oeverstraat 65
P.O.Box 1034
6801 MG Arnhem
NETHERLANDS

The book contains articles based on the papers presented at an east-west conference on language testing, which was organized at the University of Jyväskylä in 1991. The conference brought together leading experts in language testing. Their contributions address the state of language testing, policy and methodological aspects of language testing, and several empirical questions in language testing.

This publication can be obtained from
Institute for Educational Research
University of Jyväskylä
PL 35, 40351 Jyväskylä, Finland
Phone +358-41-603221
Fax +358-41-603201

Jyväskylä 1993 ISBN 951-34-0059-X

253